COMPARAÇÃO DE MÉTODOS NA ESTIMAÇÃO DO VICIO DA RAZÃO DE ERRO APARENTE.

Valdério Anselmo Reisen

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA DA COMPUTAÇÃO

CAMPINAS - SÃO PAULO
BRASIL

TÍTULO DA TESE

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. VALDÉRIO ANSELMO REISEN e aprovada pela Comissão Julgadora.

Campinas, 28 de agosto de 1987.

Prof. Dr. _____

Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do título de Mestre em Estatística.

AGRADECIMENTOS

Ao meu orientador Prof. Dr. Ademir José Petenate pela orientação, críticas e sugestões na orientação do trabalho.

Aos meus professores durante o curso de graduação na UFES, Paulo Sérgio Moala e Dirley M. dos Santos, atualmente colegas de trabalho, pelo incentivo dado para realização do curso de pós-graduação.

Aos colegas do departamento de estatística da UFES, que possibilitaram minha saída para realização do trabalho de tese.

As minhas colegas, pelo carinho e amizade que elas puderam me proporcionar, Zélia Judith Loss e Rosa Maria V. Pessoa.

Aos colegas do curso de pós-graduação em estatística na Unicamp, que direta ou indiretamente colaboraram para o desenvolvimento desse trabalho.

Aos meus amigos de São Paulo, que me proporcionaram ótimos momentos de descontração.

Aos meus professores do curso de pós-graduação do IMECC que contribuiram para o meu aprimoramento em estatística.

À Edi, pelo trabalho de datilografia.

À CAPES, CNPq e FAPESP pelo auxílio financeiro que me foi concedido para a realização do curso de pós-graduação.

ÍNDICE

pág.

CAPÍTULO 1

Teorias e Regras de Classificação

CAPÍTULO 2

Métodos de Estimação das Probabilidades de Má-classificação

pág.

CAPÍTULO 3

Comparação de Métodos na Estimação do Vício do Estimador da Razão de Erro Real

CAPÍTULO 4

Conclusão dos Resultados Experimentais *pág.*

SUMÁRIO

Na teoria de discriminação e classificação, a Razão de Erro Real associada a uma função de classificação, é dada pelas probabilidades de má classificação, e é usada para estudar o desempenho da função de classificação. Para o cálculo da Razão de Erro Real é necessário o conhecimento das distribuições populacionais que estão envolvidas no problema.

Existe uma vasta literatura à respeito da estimação da Razão de Erro Real. Em nosso trabalho, nós utilizaremos o estimador Razão de Erro Aparente que é um método comumente usado em outros trabalhos. Estudos mostram que este substima a Razão de Erro Real.

O objetivo de nosso trabalho é estudar o vício do estimador Razão de Erro Aparente, num contexto de comparar três métodos de estimação, sendo que dois métodos são não paramétricos, o "bootstrap" proposto por Efron e o "cross-validation". O terceiro é um método paramétrico proposto por Mclachlan.

A comparação foi feita sob condição do conhecimento das distribuições populacionais. Amostras de duas populações normais multivariadas com diferentes vetores de médias e matrizes de covariâncias iguais à identidade, foram geradas por processo de simulação de Monte Carlo.

A conclusão de nossos estudos é que, o método "bootstrap" mostrou-se eficiente na estimação do vício da Razão de Erro Aparente tanto quanto o estimador paramétrico, sabendo que, este último é condicionado ao conhecimento das distribuições populacionais para a sua validade.

Este trabalho está dividido em quatro capítulos. A teoria de classificação e discriminação, utilizando o método de Fisher para a construção de regras de classificação, é vista no capítulo 1. O capítulo 2 apresenta uma pesquisa bibliográfica com os métodos de estimação da probabilidade de má classificação. O procedimento de vários métodos assim como o método R, usado para obter o estimador da Razão de Erro Aparente, também são desenvolvidos.

A comparação feita entre os métodos de estimação do vício da Razão de Erro Aparente é vista no capítulo 3, com a descrição dos métodos "bootstrap", "cross-validation" e o "paramétrico" e também a metodologia de nosso trabalho e os resultados experimentais.

Finalmente, no capítulo 4, apresenta-se as conclusões do nosso trabalho.

CAPÍTULO 1

TEORIAS E REGRAS DE CLASSIFICAÇÃO

1.1 - INTRODUÇÃO:

Este capítulo tem por objetivo apresentar técnicas e regras no procedimento de discriminação e classificação. A teoria de análise discriminante de dados consiste em separar populações ou classificar novas observações em populações previamente definidas.

O problema da classificação aparece quando desejamos, de acordo com o número de medidas de um indivíduo (ou objeto), classificá-lo em uma das várias categorias (ou populações) envolvidas. Nós não podemos assumir diretamente a que categoria ele pertence mas devemos fazer uso de suas medidas (características) para classificá-lo.

A classificação de uma observação (ou objeto) em uma das populações não é feita sem erro e, portanto, serão estudadas aqui técnicas com intuito de conhecer e minimizar o erro de classificação.

Nós assumiremos um número finito de populações sendo que cada população é caracterizada pela distribuição de probabilidade de suas medidas e uma observação é um valor amostral de uma dessas populações.

Em algumas situações as distribuições de probabilidade das populações são consideradas conhecidas de início, em outras situações a forma da distribuição de probabilidade das populações é conhecida mas não conhecemos seus parâmetros e, então, fazemos o uso das amostras dessas populações para estimar as quantidades desconhecidas.

As técnicas aqui estudadas serão aplicadas para populações com distribuição normal. Em alguns resultados apresentados não nos preocupamos em fazer a demonstração, mas é dada a referência bibliográfica.

A secção 1.2 apresenta o problema de classificação e discriminação envolvendo duas populações pela função discriminante de Fisher. O problema de classificação que é o erro, e as regras que o minimizam são apresentadas na secção 1.3 e 1.4. Nesta última é feita também uma pequena abordagem da construção das regras de classificação através do procedimento de Bayes. Já a secção 1.5 faz a aplicação dos resultados das secções anteriores em populações com distribuição normal. É apresentado também no final desta secção uma rápida idéia de estimadores que independem da forma da população de origem para a Razão de Erro Real, que é uma das medidas usadas para se conhecer o erro de classificação. Finalmente, na secção 1.6, as regras são extendidas para o caso mais geral, com várias populações e aplicando-as em populações com distribuição normal.

1.2 - MÉTODO DE FISHER PARA O PROBLEMA DE CLASSIFICAÇÃO E SEPARAÇÃO (DISCRIMINAÇÃO) ENVOLVENDO DUAS POPULAÇÕES.

Aqui estaremos interessados em situações onde desejamos separar duas classes de objetos (observações) ou alocar novo objeto (observação) em uma das duas classes (ou ambas). Consideramos Π_1 e Π_2 as duas classes de objetos (observações) e $X_i' = [X_1, X_2, \ldots, X_p]$ para $i=1,2$ um vetor aletório em $\mathbb{R}^p$ o qual é usado para classificar ou separar com base em suas medidas. Os valores observados de X_i diferem de alguma forma de uma classe a outra. Através das medidas de X_i, nós construiremos uma regra para classificar uma nova observação $x \in \mathbb{R}^p$ em uma das duas classes (populações) Π_1 e Π_2. Seja $f_i(x)$ função de densidade de probabilidade associada a Π_i; $i = 1,2$. O trabalho de Fisher consiste em transformar a observação multivariada x em uma observação univariada y através de uma combinação linear, ficando assim mais fácil de trabalhar matemáticamente.

Tomamos μ_{1Y} como sendo a média dos Y's obtidos dos X's pertencentes a Π_1 e, μ_{2Y} a média dos Y's obtidos dos X's pertencentes a Π_2.

Definimos:

$\mu_1 = E(X/\Pi_1)$ = valor esperado da variável aleatória X pertencente a Π_1

$\mu_2 = E(X/\Pi_2)$ = valor esperado da variável aleatória X pertencente a Π_2

e $\Sigma_i = E(X-\mu_i)(X-\mu_i)'$, matriz de variâncias e covariâncias de Π_i onde $i=1,2$, $\Sigma_1 = \Sigma_2 = \Sigma$

Consideremos a combinação linear:

$$Y = t' X$$

através das propriedades das esperanças temos,

$$\mu_{1Y} = E(t'X/ \Pi_1) = t' \mu_1$$

$$\mu_{2Y} = E(t'X/ \Pi_2) = t' \mu_2$$

e, $$V(Y) = Var(t'X) = t'\Sigma t$$

A idéia de Fisher é selecionar a combinação linear que maximiza a distância ao quadrado entre μ_{1Y} e μ_{2Y} relativo à variância dos Y's.

$$(1) \qquad \frac{\text{(distância ao quadrado das médias dos Y's)}}{\text{(variância de Y)}} = \frac{(\mu_{1Y} - \mu_{2Y})^2}{V(Y)}$$

$$\frac{(t'\mu_1 - t'\mu_2)^2}{t'\Sigma t} = \frac{t'(\mu_1 - \mu_2)(\mu_1 - \mu_2)'t}{t'\Sigma t} = \frac{(t'\delta)^2}{t'\Sigma t}$$

onde, $$\delta = (\mu_1 - \mu_2)$$

A escolha do vetor $t' = (t_1, t_2, \ldots, t_p)$ da combinação linear que maximiza a expressão (1) é obtido a partir do seguinte resultado:

TEOREMA 1.1

Seja $\delta = \mu_1 - \mu_2$ e $Y = t'X$, então

(2) $$\frac{\text{(distância ao quadrado das médias dos Y's)}}{\text{(variância de Y)}} = \frac{(t'\delta)^2}{t'\Sigma t}$$

é maximizado pela escolha de $t = c\Sigma^{-1}\delta = c\Sigma^{-1}(\mu_1 - \mu_2)$ $\forall$ $c \neq 0$. Tomando $c = 1$ temos a função discriminante linear de Fisher,

(3) $$Y = t'X = (\mu_1 - \mu_2)'\Sigma^{-1} X$$

e o $$\max_{t \neq 0} \frac{(t'\delta)^2}{t'\Sigma t} = \delta'\Sigma^{-1}\delta$$

Prova: Johnson [1982]

Usaremos a função discriminante linear de Fisher para construir a regra de classificação. Seja $y_0 = (\mu_1 - \mu_2)'\Sigma^{-1}x_0$ a função discriminante para uma observação $x_0 \in \mathbb{R}^p$.

Calculando o ponto médio entre as populações univariadas obtemos:

$$m = 1/2\,(\mu_1 Y + \mu_2 Y) = 1/2\,(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2).$$

É fácil provar quo as desigualdades abaixo são verdadeiras,

$$E(Y/\Pi_1) - m \geq 0$$

e, $E(Y/\ \Pi_2) - m < 0$.

Aplicando esse resultado a uma observação Y_0 obtemos:

$$E(Y_0/\ \Pi_1) - m \geq 0$$

(4)

e, $E(Y_0/\ \Pi_2) - m < 0$

isto é, se X_0 pertence à Π_1, é esperado que Y_0 seja maior que o ponto médio, assim se X_0 é pertencente à Π_2 é esperado que Y_0 seja menor que o ponto médio.

Através de (4) construiremos a seguinte regra de classificação para uma observação x_0. Alocamos x_0 para Π_1 se,

(5) $$y_0 - m \geq 0 \quad \Rightarrow \quad (\mu_1 - \mu_2)'\Sigma^{-1}x_0 \geq m$$

caso contrário alocamos x_0 para Π_2.

Como os valores populacionais μ_1, μ_2 e Σ são raramente conhecidos, trataremos de estimá-los através de amostras populacionais.

Iniciaremos considerando duas amostras do vetor de observação $X'_{i(p}$ = [X_{i1}, X_{i2}, ..., X_{ip}] para i=1,2 de tamanho n_1 da população Π_1 e n_2 da população Π_2. Temos então as matrizes de observação:

$$X_1 = [x_{11}, x_{12}, \ldots, x_{1n_1}] : p \times n_1$$

$$X_2 = [x_{21}, x_{22}, \ldots, x_{2n_2}] : p \times n_2$$

Calculamos os vetores de médias e as matrizes de variâncias e covariâncias amostrais,

$$\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j} \; ; \quad \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j} : p \times 1$$

$$S_1 = \frac{1}{(n_1-1)} \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)' : p \times p$$

$$S_2 = \frac{1}{(n_2-1)} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)' : p \times p$$

Considerando $\Sigma_1 = \Sigma_2 = \Sigma$, S_1 e S_2 são ambos estimadores de Σ , consequentemente nós podemos combinar S_1 e S_2 para encontrar um estimador não viciado de Σ , considerando que as matrizes de observações X_1 e X_2 são amostras das populações Π_1 e Π_2 respectivamente.

(6)
$$S_{p \times p} = \frac{(n_1-1)}{(n_1-1) + (n_2-1)} S_1 + \frac{(n_2-1)}{(n_1-1) + (n_2-1)} S_2 =$$

$$= \frac{(n_1-1)S_1 + (n_2-1)S_2}{(n_1 + n_2 - 2)}$$

S é o estimador combinado de Σ . Substituindo μ_1, μ_2 e Σ em (3) pelas quantidades $\bar{x}_1, \bar{x}_2$ e S obtemos a **Função Discriminante Linear Amostral de Fisher**,

$$(7) \qquad y = \hat{t}'x = (\bar{x}_1 - \bar{x}_2)'S^{-1}x$$

O ponto médio $\hat{m}$, entre as duas médias amostrais das populações univariadas, $\bar{y}_1 = \hat{t}'\bar{x}_1$ e $\bar{y}_2 = \hat{t}'\bar{x}_2$ é dado por:

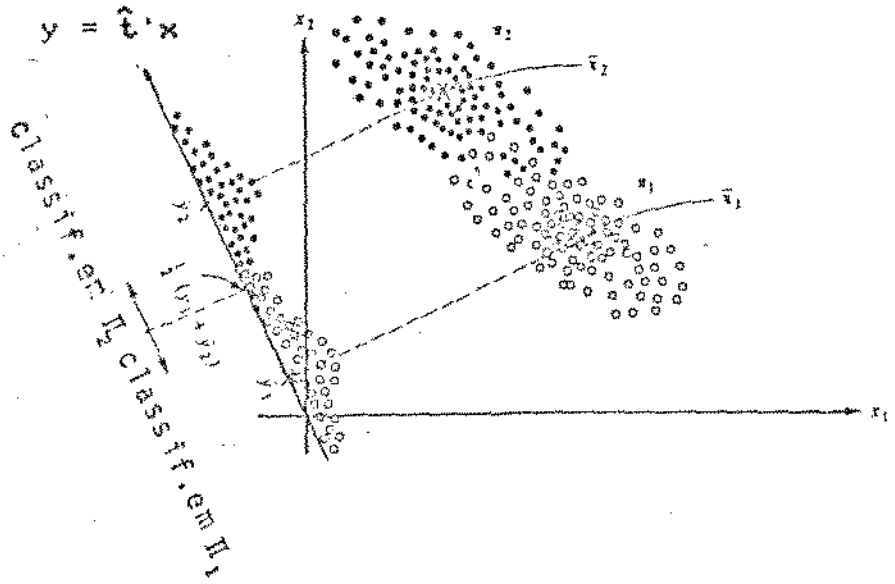
$$(8) \qquad \hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 + \bar{x}_2)$$

Construimos então, a seguinte **Regra de Classificação Amostral**. Alocamos x_0 para Π_1 se

$$(9) \qquad y_0 = (\bar{x}_1 - \bar{x}_2)'S^{-1}x_0 \geq \hat{m}$$

caso contrário alocamos x_0 a Π_2.

REPRESENTAÇÃO GRÁFICA DA FUNÇÃO DISCRIMINANTE LINEAR AMOSTRAL DE FISCHER PARA p=2



De acordo com o Teorema 1.1 , obtemos o seguinte resultado para a expresssão (7):

TEOREMA 1.2

Seja $d = (\bar{x}_1 - \bar{x}_2)$, a combinação linear $y = \hat{t}'x$ onde, $\hat{t} = S^{-1}(\bar{x}_1 - \bar{x}_2)$ maximiza a expressão,

$$\frac{\text{(distância ao quadrado das médias amostrais dos y's)}}{\text{(variância amostral de y)}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s^2_y}$$

$$\frac{(\hat{t}'\bar{x}_1 - \hat{t}'\bar{x}_2)^2}{\hat{t}'S\,\hat{t}} = \frac{(\hat{t}'d)^2}{\hat{t}'S\,\hat{t}}$$

e o

$$\text{(10)} \qquad \max_{\hat{t}\neq 0} \frac{(\hat{t}'d)^2}{\hat{t}'S\,\hat{t}} = d'S^{-1}d = (\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) = D^2$$

onde D^2 é a distância amostral ao quadrado entre duas amostras de duas populações.

A prova do teorema 1.2 é equivalente ao teorema 1.1.

1.3 - O PROBLEMA EM CLASSIFICAÇÃO

Nos problemas de classificação estamos sujeitos aos possíveis erros de classificação, isto é, alocar uma observação a certa população quando na realidade esta observação pertence a outra população. Isto pode ocorrer por vários fatores e um deles é devido a que as características medidas nas duas populações possam ser bastante semelhantes, facilitando assim a ocorrência da má classificação.

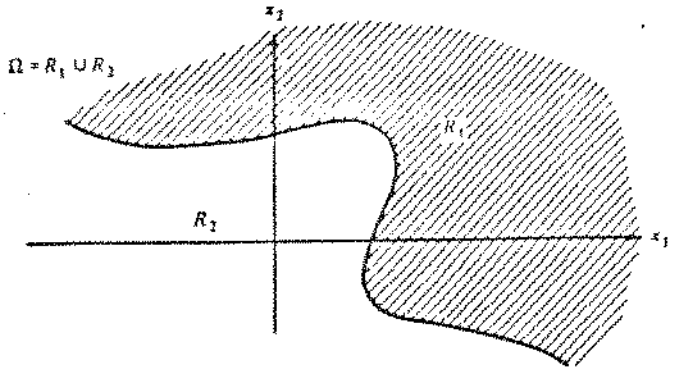
Na construção do procedimento de classificação é desejado que se minimize a probabilidade de má classificação, ou melhor, é desejado minimizar o erro de classificação.

Outro aspecto que nós abordaremos aqui no problema de classificação é o custo. Suponha que classificar um objeto em Π_1 sendo ele pertencente a Π_2 representa um erro mais sério com respeito ao custo do que classificar um objeto em Π_2 sendo pertencente a Π_1. Então teremos que ser mais cautelosos na forma de classificação.

Nós iniciaremos agora o procedimento do problema da classificação envolvendo duas classes (populações). Mais adiante trataremos do caso mais geral.

As regras de classificação ou alocação são construídas através das informações amostrais, então, necessitamos de amostras aleatórias com medidas características de cada população Π_1, Π_2. Dividimos Ω, o espaço amostral em duas regiões R_1, R_2, de tal forma que se uma nova observação estiver em R_1 é alocada a Π_1. Da mesma forma, se estiver em R_2, é alocada a Π_2. Podemos então escrever $R_1 = \Omega - R_2$; (R_1 e R_2 são mutualmente exclusivos e exaustivos).

Para p = 2 é feita a representação gráfica abaixo, para as regiões R_1 e R_2 de classificação:



Sejam $f_1(x)$ e $f_2(x)$ funções de densidade de probabilidade associadas ao vetor **px1** aleatório da variável **X** para as populações Π_1 , Π_2 respectivamente.

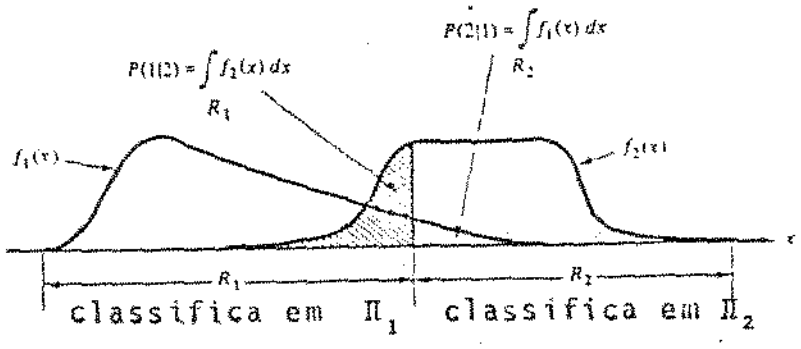
A probabilidade condicional, p(2/1) de classificar uma observação a Π_2, sendo pertencente a Π_1 é:

$$(11) \qquad p(2/1) = p(x \in R_2 / \Pi_1) = \int_{R_2} f_1(x)dx$$

de mesma forma a probabilidade condicional p(1/2) de classificar uma observação em Π_1 quando é pertencente a Π_2 é:

$$(12) \qquad p(1/2) = p(x \in R_1 / \Pi_2) = \int_{R_1} f_2(x)dx$$

A figura abaixo ilustra essas duas probabilidades para o caso em que p = 1.



APRESENTAÇÃO DAS PROBABILIDADES DE CLASSIFICAÇÃO

Definimos os seguintes eventos para o cálculo das probabilidades de má classificação:

A = observação pertence a Π_1

B = observação é corretamente classificada

$\bar{A}$ = observação pertence a Π_2

$\bar{B}$ = observação não é corretamente classificada

p_1 e p_2 são probabilidades a priori das populações Π_1 e Π_2 respectivamente.

(13) $p(\text{corretamente classificada a } \Pi_1) = p(A \cap B) =$

$= p(B/A)\ p(A) = p(x \in R_1/\ \Pi_1)p_1$

(14) $p(\text{má classificada em } \Pi_1) = p(\bar{A} \cap \bar{B}) =$

$= p(\bar{B}/\bar{A})\ p(\bar{A}) = p(x \in R_1/\ \Pi_2)p_2$

(15) $p(\text{corretamente classificada em } \Pi_2) = p(\bar{A} \cap B) =$

$= p(B/\bar{A})\ p(\bar{A}) = p(x \in R_2/\ \Pi_2)p_2$

(16) $p(\text{má classificada em } \Pi_2) = p(A \cap \bar{B}) =$

$= p(\bar{B}/A)\ p(A) = p(x \in R_2/\ \Pi_1)p_1$

É interessante no problema de classificação incluir o custo de má classificação pois, mesmo sendo pequeno o erro de má classificação, pode acontecer que o custo desse erro seja alto.

Definimos então a matriz dos custos:

		DECISÃO ESTATÍSTICA	
		Π_1	Π_2
VERDADEIRA POPULAÇÃO	Π_1	0	c(2/1)
	Π_2	c(1/2)	0

Os custos são zeros para as classificações corretas e c(2/1) quando a observação é incorretamente classificada em Π_2 assim como c(1/2) quando a observação é incorretamente classificada em Π_1.

O procedimento aqui abordado será aquele que minimiza o custo esperado de má classificação (CEM), que é definido como o produto dos elementos da diagonal da matriz acima com suas respectivas probabilidades. Temos então,

(17) $CEM = c(2/1)\ p(2/1)\ p_1 + c(1/2)\ p(1/2)p_2$

1.4 - REGRAS DE CLASSIFICAÇXO ENVOLVENDO DUAS POPULAÇÕES

Aqui apresentaremos regras de classificação baseadas no CEM, onde desejamos minimizá-lo. De início trataremos de construir as regras envolvendo duas populações, o caso mais geral será tratado no final deste capítulo.

A idéia é encontrar regiões R_1, R_2 que minimizam o CEM. Segue então o primeiro resultado:

TEOREMA 1.3

As regiões R_1, R_2 que minimizam o CEM (17) são definidas para os valores de x que satisfazem as seguintes desigualdades:

(18)
$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{c(1/2)\ p_2}{c(2/1)\ p_1}$$

$$R_2: \frac{f_1(x)}{f_2(x)} < \frac{c(1/2)\ p_2}{c(2/1)\ p_1}$$

Prova: Jonhson [1982]

CASOS ESPECIAIS DO TEOREMA 1.3

(a) $p_2 = p_1$ (iguais probabilidades à priori)

$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{c(1/2)}{c(2/1)}$$

$$R_2: \frac{f_1(x)}{f_2(x)} < \frac{c(1/2)}{c(2/1)}$$

(b) (c (1/2) = c (2/1), custos iguais)

(19)
$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}$$

$$R_2: \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$$

(c) ocorrer (a) e (b) conjuntamente

$$R_1: \frac{f_1(x)}{f_2(x)} \geq 1$$

$$R_2: \frac{f_1(x)}{f_2(x)} < 1$$

São três os casos possíveis no uso do teorema 1.3. Podemos não conhecer as probabilidades p_1, p_2 e então consideramo-las iguais, ítem (a). De mesma forma, os custos não sendo conhecidos, usamos o ítem (b) e finalmente, nas duas situações acima, isto é, não conhecendo p_1, p_2, c(1/2), c(2/1), adotamos o ítem (c).

Agora faremos uma apresentação dos estudos de Anderson [1958].

As regras de classificação são baseadas no Teorema de Bayes. Os resultados são aqui apresentados de acordo com as definições dadas no início deste capítulo.

Sejam as probabilidades a posteriori:

$p(\Pi_1/x)$ = probabilidade a posteriori associada a Π_1,

$p(\Pi_2/x)$ = probabilidade a posteriori associada a Π_2

Usando o teorema da probabilidade total obtemos:

$$p(\text{observar } x) = p(\text{observar } x/\Pi_1)p_1 + p(\text{observar } x/\Pi_2)p_2.$$

Pela Regra de Bayes obtemos $p(\Pi_1/x)$ e $p(\Pi_2/x)$.

$$p(\Pi_1/x) = \frac{p(\Pi_1 \text{ ocorrer e observar } x)}{p(\text{observar } x)}$$

$$= \frac{p(\text{observar } x/\ \Pi_1)p_1}{p(\text{observar } x/\ \Pi_1)p_1 + p(\text{observar } x/\ \Pi_2)p_2} =$$

$$= \frac{f_1(x)p_1}{f_1(x)p_1 + f_2(x)p_2}$$

de mesma forma,

$$p(\ \Pi_2/x) = \frac{f_2(x)p_2}{p_2f_2(x) + p_1f_1(x)}$$

Através dessas probabilidades, encontraremos regras que minimizam o CEM (17).

Seja c(1/2) = c(2/1), então a expressão (17) é a probabilidade total de má classificação, assim escrita,

$$PTM = p_1p(2/1) + p_2p(1/2).$$

Para um valor observado x, minimizaremos a probabilidade total de má-classificação classificando-o na população que apresenta maior probabilidade a posteriori. Se,

$$\frac{f_1(x)p_1}{f_1(x)p_1 + f_2(x)p_2} \geq \frac{f_2(x)p_2}{f_1(x)p_1 + f_2(x)p_2}$$

nós classificamos x em Π_1, caso contrário em Π_2.

Sendo assim, minimizaremos a probabilidade de má classificação em cada ponto x observado e, então, minimizaremos em todo o espaço. Segue então a regra:

As regiões R_1, R_2 que minimizam a probabilidade total de má classificação são definidas da forma:

$$R_1: p_1 f_1(x) \geq p_2 f_2(x)$$

$$R_2: p_1 f_1(x) < p_2 f_2(x)$$

Vemos que este resultado já é conhecido, isto é, equivale às regras vistas anteriormente.

A demonstração do resultado acima e uma abordagem mais geral deste procedimento é mostrado por Anderson [1958]. Muitos de seus resultados são equivalentes aos aqui apresentados.

1.5 - REGRAS DE CLASSIFICAÇÃO COM DUAS POPULAÇÕES NORMAIS MULTIVARIADAS

Aqui agora usaremos as regras de classificação desenvolvidas nas secções anteriores considerando as populações Π_1 e Π_2 normais com densidades $f_1(x)$ e $f_2(x)$, isto é, Π_i é distribuído $N(\mu_i, \Sigma_i)$, $i = 1,2$.

1.5.1 - 1º Caso:

$$\Sigma_1 = \Sigma_2 = \Sigma$$

Sejam μ_i, Σ conhecidos e,

$$(20) \qquad f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-1/2\,(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right]$$

com $i = 1,2$; a função de densidade de $X' = [X_1, X_2, \ldots, X_p]$, para as populações Π_1 e Π_2. Usando o teorema 1.3 podemos encontrar as regiões R_1, R_2 que minimizam o CEM (17).

Portanto,

(21)

$$R_1: \exp\left[-1/2\,(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + 1/2\,(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right] \geq$$

$$\geq \frac{c(1/2)p_2}{c(2/1)p_1}$$

R_2: $\exp[-1/2\ (x-\mu_1)'\Sigma^{-1}(x-\mu_1) + 1/2\ (x-\mu_2)'\Sigma^{-1}(x-\mu_2)] <$

$$< \frac{c(1/2)p_2}{c(2/1)p_1}$$

De acordo com as regiões acima segue o seguinte teorema:

TEOREMA 1.4

Sejam Π_1, Π_2 populações com densidades dadas em (20) e x_0 uma observação no espaço p-dimensional. A regra de classificação que minimiza o CEM é dada por:

classifica x_0 em Π_1 se,

$$(22)\ (\mu_1-\mu_2)'\Sigma^{-1}x_0 - 1/2\ (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geq \ln \frac{c(1/2)p_2}{c(2/1)p_1}$$

caso contrário x_0 é classificado em Π_2.

Prova: usando as expressões de (21) e um simples algebrismo, concluímos a demonstração do teorema.

Observamos que o resultado do teorema 1.4 para o caso em que $c(1/2)p_1/c(2/1)p_2 = 1$ é equivalente à expressão (5) que utiliza a **Função Discriminante Linear de Fisher.**

No resultado anterior consideramos que os parâmetros μ_1, μ_2, Σ são conhecidos. Geralmente isto não acontece. Neste caso, utilizamos amostras das populações para estimar as quantidades μ_1, μ_2, Σ . Temos então, a seguinte regra de classificação amostral, que estima o mínimo do CEM:

classificamos x_0 em Π_1 se,

$$(23) \qquad (\bar{x}_1 - \bar{x}_2)' S^{-1} x_0 - 1/2 \ (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2) \geq \ln \frac{c(1/2)p_2}{c(2/1)p_1}$$

caso contrário classificamos x_0 em Π_2.

O primeiro termo dessa expressão é exatamente a **Função Linear Amostral de Fisher** (7). A expressão,

$$(\bar{x}_1 - \bar{x}_2)' S^{-1} x_0 - 1/2 \ (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)$$

é conhecida como **Função de Classificação de Anderson.**

Anderson [1958] apresenta o seguinte teorema para a função acima:

TEOREMA 1.5

Seja $W = (\bar{x}_1 - \bar{x}_2)' S^{-1} x - 1/2 \ (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)$

com, $\bar{x}_1$ = média da amostra de tamanho n_1 de $\Pi_1 \sim N(\mu_1, \Sigma)$

$\bar{x}_2$ = média da amostra de tamanho n_2 de $\Pi_2 \sim N(\mu_2, \Sigma)$

e, S estimador de Σ .

A distribuição de W com $n_1 \to \infty$ e $n_2 \to \infty$ é ,

$N(1/2\ \Delta^2, \Delta^2)$ se x é distribuído de acordo com Π_1

e é $N(-1/2\Delta^2, \Delta^2)$ se x é distribuído de acordo com Π_2.

$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$; distância ao quadrado de Mahalanobis entre duas populações.

Prova: Ver Anderson [1958].

Para encontrar a expressão que minimiza o CEM foi assumido que as populações Π_1, Π_2 eram normais multivariadas com densidade $f_i(x)$; $i = 1,2$ e; as quantidades μ_1, μ_2, Σ conhecidas. A expressão (23) é uma simples estimativa da expressão (22) que minimiza o CEM. Não podemos esperar que um resultado particular da expressão (23) possa minimizar o CEM mas, para tamanhos grandes das amostras é razoável esperar um resultado satisfatório.

1.5.2 - 2º Caso:

CONSTRUÇÃO DA REGRA DE CLASSIFICAÇÃO QUANDO $\Sigma_1 \neq \Sigma_2$

Os procedimentos vistos até agora foram obtidos considerando matrizes de variâncias e covariâncias iguais para as duas populações Π_1 e Π_2. Agora desenvolveremos regras considerando $\Sigma_1 \neq \Sigma_2$

Nós vimos que as regiões R_1, R_2 que minimizam o custo esperado de má classificação são encontradas de acordo com a razão das densidades, teorema 1.3,

$$\frac{f_1(x)}{f_2(x)} = \frac{(2\pi)^{p/2} |\Sigma_2|^{1/2} \exp[-1/2(x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)]}{(2\pi)^{p/2} |\Sigma_1|^{1/2} \exp[-1/2(x - \mu_2)' \Sigma_2^{-1} (x - \mu_2)]} =$$

(24)

$$= \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} \exp[-1/2(x - \mu_1)'\Sigma_1^{-1}(x - \mu_1) + 1/2(x - \mu_2)'\Sigma_2^{-1}(x - \mu_2)] = b$$

Pelo teorema 1.3,

$$R_1 : b \geq \frac{c(1/2)p_2}{c(2/1)p_1}$$

Aplicando logaritmo:

$$(25) \quad \ln \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} - 1/2(x - \mu_1)'\Sigma_1^{-1}(x - \mu_1) +$$

$$+ 1/2(x - \mu_2)'\Sigma_2^{-1}(x - \mu_2) \geq \ln \frac{c(1/2)p_2}{c(2/1)p_1} ,$$

$$\ln \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} - 1/2x'(\Sigma_1^{-1} - \Sigma_2^{-1})'x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x$$

$$- 1/2(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) \geq \ln \frac{c(1/2)p_2}{c(2/1)p_1} ,$$

toma-se $K = \ln \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} - 1/2(\mu'_1\Sigma_1^{-1}\mu_1 - \mu'_2\Sigma_2^{-1}\mu_2)$;

temos então o resultado:

TEOREMA 1.6

As regiões que minimizam o CEM são dadas pelos valores de x que satisfazem as seguintes desigualdades:

(26)

$$R_1: -1/2\, x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1})x + K \geq \ln \frac{c(1/2)p_2}{c(2/1)p_1}$$

$$R_2: -1/2\, x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1})x + K < \ln \frac{c(1/2)p_2}{c(2/1)p_1}$$

Para o caso em que $\Sigma_1 = \Sigma_2$, as regiões são equivalentes ao resultado do teorema 1.4. As regiões de (26) são definidas pela forma quadrática de x.

O resultado acima depende dos parâmetros reais da população, Σ_1, Σ_2, Π_1, Π_2. Na prática, essa regra de classificação é construída com base nas amostras das populações Π_1, Π_2, e assim, substituindo as quantidades amostrais $\bar{x}_1$, $\bar{x}_2$, S_1, S_2 em μ_1, μ_2, Σ_1, Σ_2, respectivamente. As expressões para calcular $\bar{x}_1$, $\bar{x}_2$, S_1, S_2 foram dadas no início deste capítulo.

Temos então a regra de classificação quadrática amostral para populações com distribuição normal e diferentes matrizes de covariâncias e variâncias.

Classificamos uma observação em x_0 em Π_1 se,

$$(27) \quad -1/2\, x_0'(S_1^{-1} - S_2^{-1})x_0 + (\bar{x}_1' S_1^{-1} - \bar{x}_2' S_2^{-1})x_0 + \hat{K} \geq \ln \frac{c(1/2)p_2}{c(2/1)p_1}$$

caso contrário, classificamos x_0 em Π_2.

1.5.3 - AVALIAÇÃO DO PROCEDIMENTO DE CLASSIFICAÇÃO

Agora, através de uma aplicação considerando as populações normais com suas quantidades μ_1, μ_2, Σ conhecidas, avallaremos as regras de classificação.

Seja c(2/1) = c(1/2) e então o custo de má classificação é o valor correspondente da probabilidade total de má classificação PTM,

(28) $$PTM = p_1 p(2/1) + p(1/2) p_2 =$$

$$= p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

para qualquer escolha das regiões R_1, R_2. O menor valor da expressão (28), com a escolha certa de R_1, R_2 é definida por Johnson [1982] como a **Razão do Erro ótima** (REO),

(29) $$REO = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

para R_1, R_2 escolhidos da expressão (21).

De acordo com as expressões de (21); R_1, R_2 são definidas como:

$$R_1 : (\mu_1 - \mu_2)'\Sigma^{-1}x - 1/2(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln \frac{p_2}{p_1}$$

(30)

$$R_2 : (\mu_1 - \mu_2)'\Sigma^{-1} x - 1/2(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < \ln \frac{p_2}{p_1}$$

R_1 e R_2 são regiões que minimizam o PTM.

Consideramos $p_1 = p_2 = 1/2$; então,

$$\ln \frac{p_2}{p_1} = 0.$$

Escrevemos a expressão, $y = (\mu_1 - \mu_2)'\Sigma^{-1} x$

então,

$$R_1 (y) : y \geq 1/2(\mu_1 - \mu_2)'\Sigma^{-1} (\mu_1 + \mu_2)$$

$$R_2 (y) : y < 1/2(\mu_1 - \mu_2)'\Sigma^{-1} (\mu_1 + \mu_2)$$

Como Y é uma combinação linear de variáveis com distribuição normal, Y também é normal univariada. Temos então,

$$E(Y) = E((\mu_1 - \mu_2)'\Sigma^{-1} X) = (\mu_1 - \mu_2)'\Sigma^{-1} E(X)$$

Se X é pertencente a Π_1 então $E(Y) = \mu_1 Y = (\mu_1 - \mu_2)'\Sigma^{-1} \mu_1$

Se X é pertencente a Π_2 então $E(Y) = \mu_2 Y = (\mu_1 - \mu_2)'\Sigma^{-1} \mu_2$

$$V(Y) = V ((\mu_1 - \mu_2)'\Sigma^{-1} X) = (\mu_1 - \mu_2)'\Sigma^{-1}\Sigma\,\Sigma^{-1}(\mu_1 - \mu_2) =$$

$$= (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) = \Delta^2$$

Δ^2 é a distância ao quadrado de Mahalanobis. Calculando as probabilidades de má classificação, obtemos,

1) $p(2/1)$ = p(classificar em Π_2 sendo a observação pertencente a Π_1)

$$= p(Y < 1/2\,(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)) =$$

$$p\left(\frac{Y - \mu_{1Y}}{\sqrt{V(Y)}} < \frac{1/2\,(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) - \mu_{1Y}}{\sqrt{V(Y)}}\right)$$

$$p(2/1) = p\left(z < \frac{1/2(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) - (\mu_1 - \mu_2)\Sigma^{-1}\mu_1}{\Delta}\right) =$$

$$= p\left(z < \frac{1/2(\mu_1 - \mu_2)'\Sigma^{-1}[\mu_1 + \mu_2 - 2\mu_1]}{\Delta}\right) =$$

$$= p\left(z < \frac{-1/2((\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2))}{\Delta}\right) =$$

$$= p(z < -1/2\,\Delta) = G(-\Delta/2)$$

2) $p(1/2) = p(Y > 1/2(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$

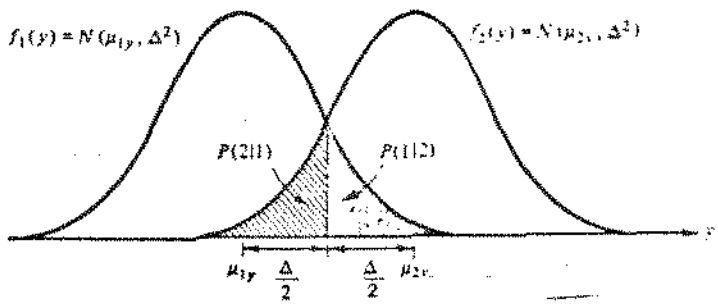
$$p\left(\frac{Y - \mu_{2Y}}{\sqrt{V(Y)}} \geq \frac{1/2(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) - \mu_{2Y}}{\Delta}\right) =$$

$$= p\left(z \geq \frac{1/2(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2 - 2\mu_2)}{\Delta}\right) =$$

$$= p\left(z \geq \frac{1/2(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)}{\Delta}\right) =$$

$= 1 - G(\Delta/2) = G(-\Delta/2)$, portanto $p(1/2) = p(2/1)$.

A figura abaixo ilustra essas probabilidades:



De acordo com os resultados acima a **Razão de Erro ótima** é então calculada,

$$REO = 1/2\ G(-\Delta/2) + 1/2\ G(-\Delta/2) = G(-\Delta/2)$$

que é o valor mínimo da PTM para um certo valor de Δ .

Se tomarmos $\Delta^2 = 2 \Rightarrow \Delta = \sqrt{2} = 1,41$; então,

$$REO = G(-1,41/2) = G(-0,707) = 0,22$$

Vemos pelo resultado, considerando valores exatos das populações, que temos uma chance de 22% de ítens que são erradamente classificados. Não temos um procedimento com resultados satisfatórios, para essa situação a probabilidade de erro é grande, por isso, devemos estar bastante atentos a este tipo de procedimento.

Trabalhando com as quantidades amostrais $\bar{x}_1$, $\bar{x}_2$, S, o desempenho da função de classificação amostral pode ser avaliado calculando a **Razão do Erro Real (RER)**, assim definida,

$$(31) \qquad RER = p_1 \int_{\hat{R}_2} f_1(x)dx + p_2 \int_{\hat{R}_1} f_2(x)dx$$

onde, $\hat{R}_1$ e $\hat{R}_2$ são as regiões de classificação amostral construídas através das amostras de tamanho n_1 e n_2 das populações, usando os resultados $\bar{x}_1$, $\bar{x}_2$ e S para estimar μ_1, μ_2 e Σ nas expressões de (30).

As regiões $\hat{R}_1$ e $\hat{R}_2$ são definidas pelos conjuntos dos x's, para os quais as seguintes desigualdades são satisfeitas,

$$\hat{R}_1: (\bar{x}_1 - \bar{x}_2)'S^{-1}x - 1/2(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln \frac{c(2/1)p_2}{c(1/2)p_1}$$

$$\hat{R}_2: (\bar{x}_1 - \bar{x}_2)'S^{-1}x - 1/2(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 + \bar{x}_2) < \ln \frac{c(2/1)p_2}{c(1/2)p_1}$$

Notamos que para se calcular o valor da RER (31) precisamos conhecer $f_i(x)$; $i = 1,2$ .Johnson faz uma pequena apresentação dos estimadores da RER em seu texto. Aqui também apresentaremos os estimadores de uma forma resumida, não questionando a eficiência de cada um.

Existe um tipo de medida da função de classificação que não depende da forma da população de origem, pode ser calculada para qualquer procedimento de classificação e é chamada **Razão de Erro Aparente (REA)**. É definida como a fração de observações das amostras que são má classificadas pela função de classificação amostral. REA é fácil de ser calculada mas tende a subestimar RER a menos que n_1 e n_2, tamanhos amostrais sejam grandes. Tomamos n_1 observações de Π_1 e n_2 observações de Π_2, construimos a função de classificação amostral e avaliamos cada observação das amostras na função. Obtemos então a seguinte matriz:

DECISÃO ESTATÍSTICA

VERDADEIRA POPULAÇÃO	Π_1	Π_2	
Π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
Π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

onde,

n_{1C} = números de ítens de Π_1 corretamente classificados em Π_1

n_{1M} = números de ítens de Π_1 mal classificados (classif. em Π_2)

n_{2C} = números de ítens de Π_2 corretamente classificados em Π_2

n_{2M} = números de ítens de Π_2 mal classificados (classif. em Π_1)

A razão de erro aparente é:

$$REA = \frac{n_1 M + n_2 M}{n_1 + n_2}$$

Um outro procedimento para estimar a razão de erro real é dividir as amostras em dois conjuntos, isto é, as n_1 observações de Π_1 é dividida em n_{1F} e n_{1T} observações, o mesmo para n_2. As amostras de n_{1F} observações são usadas para construção da função de classificação enquanto que, as amostras de n_{1T} são usadas na função para verificar quantas observações são má classificadas; a mesma idéia para as n_2 observações.

O estimador da razão de erro real é determinada pela proporção de membros mal classificados nas amostras n_{1T} e n_{2T}. Este procedimento apresenta dois principais defeitos:

1 - Requer grandes amostras.

2 - A função avaliada não é a função de interesse, não são usadas todas as observações da amostra para construção da função e nessa situação perdem-se informações importantes.

Outro procedimento é proposto por Lachenbruch's e Mickey [1968]:

1 - Iniciamos com as observações Π_1, omitimos uma observação deste grupo e construímos a função com as $n_1 - 1$ e n_2 observações;

2 - classificamos a observação que é retirada da amostra de Π_1 na função construída no ítem 1;

3 - repetimos os itens 1 e 2 até todas que as observações de Π_1 sejam classificadas;

4 - repetimos este procedimento para Π_2.

Seja n_{1M} sendo o número de observações de Π_1 mal classificados e n_{2M} o número de observações de Π_2 mal classificados. As estimativas de p(2/1) e p(1/2), probabilidade de má classificação são dadas por:

$$\hat{p}(2/1) = \frac{n_{1M}}{n_1}$$

$$\hat{p}(1/2) = \frac{n_{2M}}{n_2}$$

este procedimento é também chamado de **cross-validation**, que será apresentado mais detalhado em outro capítulo.

O estimador da Razão de Erro Real (RER) é dada por:

$$\widehat{E(RER)} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} .$$

1.6 - PROBLEMA DE CLASSIFICAÇÃO ENVOLVENDO VÁRIAS POPULAÇÕES

Nós agora consideraremos o problema de classificar uma observação em uma de várias populações.

Seja $\Pi_1, \Pi_2, \ldots, \Pi_g$; g populações com densidades $f_1(x)$, $f_2(x), \ldots, f_g(x)$; respectivamente. Desejamos dividir o espaço de observações em g regiões mutualmente exclusivas e exaustivas R_1, R_2, ..., R_g. Se a observação estiver em R_i classificaremos como sendo pertencente a Π_i. Começamos então a definir:

p_i = probabilidade a priori da população i; Π_i = 1, 2,...,g

$c(K/i)$ = custo associado ao classificar um item a Π_k

quando é pertencente a Π_i, quando $K=i$; $c(K/i)=0$.

CÁLCULO DO CUSTO ESPERADO CONDICIONAL DE MÁ CLASSIFICAÇÃO DE X PERTENCENTE A Π_1.

(33) $CEM(1) = p(2/1)\, c(2/1) + p(3/1)\, c(3/1) + \ldots + p(1/g)\, c(g/1) =$

$$= \sum_{t=2}^{g} p(t/1)\, c(t/1),$$

De mesma forma, podemos calcular CEM(i); i= 2,...,g.

Obtemos o Custo Total de Má Classificação multiplicando cada CEM com suas respectivas probabilidades e fazendo a soma geral desses fatores. Então,

(34) $$CEM = p_1 CEM(1) + p_2 CEM(2) + \ldots + p_g CEM(g)$$

$$p_1\Big[\sum_{t=2}^{g} p(t/1)\, c(t/1)\Big] + p_2\Big[\sum_{\substack{t=1 \\ t\neq 2}}^{g} p(t/2)\, c(t/2) + \ldots +$$

$$+ p_g\Big[\sum_{t=1}^{g-1} p(t/g)\, c(t/g)\Big] =$$

$$= \sum_{i=1}^{g} p_i \Big(\sum_{\substack{t=1 \\ t\neq i}}^{g} p(t/i)\, c(t/i)\Big);$$

que é o custo total de má classificação.

Um procedimento ótimo de classificação é encontrar as regiões R_1, R_2,..., R_g que minimiza o CEM (34).

TEOREMA 1.7

As regiões de classificação que minimizam o CEM (34) são definidas da seguinte forma: classificamos x em Π_k se,

(35) $$\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(x)\, c\,(K/i) \text{ é o menor valor.}$$

Prova: Anderson [1958].

Suponhamos um caso particular do teorema 1.7; considerando os custos iguais; nós classificamos x em Π_k se,

(36) $$\sum_{\substack{i=1 \\ i\neq k}}^{g} f_i(x) p_i;$$

for o menor valor. Como consequência, a expressão (36) será menor, pois omitimos o maior termo $p_k f_k(x)$, resultando assim uma simples regra de classificação.

REGRA DE CLASSIFICAÇXO QUE MINIMIZA O CEM PARA CUSTOS IGUAIS

Classificar x em Π_k se,

(37) $$p_k f_k(x) > p_i f_i(x), \forall\ i \neq k$$

ou equivalente se,

$$\ln p_k f_k(x) > \ln p_i f_i(x).$$

Notamos que este resultado é o equivalente apresentado que minimiza a probabilidade a posteriori $p(\Pi_k/x)$.

REGRA DE CLASSIFICAÇXO COM POPULAÇÕES NORMAIS

Nós agora aplicaremos a teoria anterior no caso em que temos populações com distribuição normal multivariada. Seja f_i com distribuição $N(\mu_i, \Sigma_i)$, então, a densidade de Π_i é,

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp[-1/2(x - \mu_i)' \Sigma_i^{-1}(x - \mu_i)] \quad i = 1,\ldots,g.$$

Considerando a expressão (37) classificamos x em Π_k se,

(38) $$\ln p_k f_k(x) =$$

$$= \ln p_k - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) =$$

$$= \max_i \ln p_i f_i(x)$$

Na expressão (38) podemos ignorar o termo $p/2 \ln (2\pi)$, pois é o mesmo em todas as populações.

Nós definimos a expressão de Discriminação Quadrática para a i-ésima população como sendo:

$$(39) \qquad d_i(x) = -\frac{1}{2} \ln |\Sigma_i| - \frac{1}{2} (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \ln p_i$$

$i = 1, 2 \ldots g.$

Considerando que $c(i/i) = 0$ e $c(k/i) = 1 \; \forall \; k = i$, construimos a seguinte regra que maximiza a Probabilidade Total de Má classificação (PTM): classificar x em Π_k se,

$$(40) \qquad d_k = \max (d_i(x)) \; ; \forall \; i = 1, 2, \ldots, g$$

onde, $d_i(x)$ é dado por (39).

Um caso particular da expressão (39) é quando consideramos que $\Sigma_i = \Sigma_j \; ; \forall \; i, j = 1, 2, \ldots, g$. Obtemos então nesse caso,

$$(41) \qquad d_i(x) = \mu_i' \Sigma^{-1} x - 1/2 \; \mu_i' \Sigma^{-1} \mu_i + \ln p_i$$

e a regra de classificação é a mesma vista anteriormente em (40).

Em muitas situações na prática μ_i, Σ_i são desconhecidos e então usamos as amostras para obtermos as estimativas. Tomando uma amostra de tamanho n_i da população Π_i, as quantidades amostrais são:

$\bar{x}_i$ = vetor de média amostral

S_i = matriz de covariâncias amostrais.

O estimador da expressão de discriminação quadrática (39) é escrita na forma,

$$(42) \quad \hat{d}_i(x) = -1/2 \ln |S_i| - 1/2 (x - \bar{x}_i)' S_i^{-1} (x - \bar{x}_i) + \ln p_i \ .$$

Temos então a regra de classificação amostral. Classificamos x em Π_k se,

$$\hat{d}_k(x) = \max (\hat{d}_i(x)) \ ; \forall \ k, i = 1, 2, \ldots, g.$$

Considerando novamente que $\Sigma_i = \Sigma_j \ ; \forall \ i, j$; estimamos a expressão (41) da seguinte forma:

$$\hat{d}_i(x) = \bar{x}'_i S^{-1} x \ - 1/2 \ \bar{x}'_i S^{-1} \bar{x}_i + \ln p_i \ ; \text{ onde,}$$

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \ldots + (n_g - 1)S_g}{(n_1 + n_2 + \ldots + n_g - g)}$$

$$S_i = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' ; \ i = 1, 2 \ldots g$$

CAPÍTULO 2

MÉTODOS DE ESTIMAÇÃO DAS PROBABILIDADES DE MÁ CLASSIFICAÇÃO

2.1 - INTRODUÇÃO

Vários métodos de estimação da probabilidade de má classificação em análise discriminante são aqui apresentados. Anderson [1958, 1972] e Okamoto [1963], desenvolveram assintóticamente as probabilidades não condicionais de má classificação baseada na função de Anderson vista no capítulo 1, como serão apresentadas na secção 2.2 deste capítulo.

Amostras normais de populações multivariadas foram geradas através de ensaios de Monte Carlo para avaliar, com as verdadeiras probabilidades condicionais de má classificação, o desempenho dos métodos de estimação dessas probabilidades, métodos ditos empíricos e não empíricos para diferentes combinações de amostras e diferentes números de parâmetros. Estes são apresentados na secção 2.3 de acordo com o trabalho de Lachenbruch e Mickey [1968]. A conclusão de seus resultados é que os métodos mais usados são relativamente pobres em relação aos novos por eles propostos.

No final desta secção é feita uma pequena referência do trabalho de Sorum [1971], onde compara os métodos propostos pelos autores anteriores através do Erro Quadrático Médio Assintótico, propondo também outros métodos de estimação envolvendo amostras testes, para o caso de populações univariadas e multivariadas.

Algumas de suas conclusões estão em concordância com o trabalho proposto por Lachenbruch e Mickey.

2.2 - EXPANSXO ASSINTÓTICA DAS PROBABILIDADES DE MA-CLASSIFICAÇXO

Seja x um vetor aleatório de observações no espaço p-dimensional de uma das duas populações normais multivariadas Π_1 ou Π_2 com vetores médias μ_1 ou μ_2 e matriz de covariâncias e variâncias Σ .

Suponhamos que exista uma amostra de n_1 observações multivariadas de Π_1 e n2 observações multivariadas de Π_2, obtendo assim os estimadores $\bar{x}_1$, $\bar{x}_2$, S de μ_1, μ_2, Σ respectivamente.

A regra de classificação aqui usada para o cálculo das probabilidades de má classificação foi desenvolvida no capítulo 1, que é usualmente chamada de Função de Classificação de Anderson, assim definida:

$$W(x) = (\bar{x}_1 - \bar{x}_2)'S^{-1}x - 1/2(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 + \bar{x}_2) =$$

$$= [x - 1/2(\bar{x}_1 + \bar{x}_2)]'S^{-1}(\bar{x}_1 - \bar{x}_2)$$

onde a primeira parte de W(x) é a **Função Discriminante Linear Amostral de Fisher.**

Os resultados aqui apresentados foram desenvolvidos por Anderson [1958,1972], Okamoto [1963], onde este obtem expansão assintótica da distribuição de W no termo de N^2 onde, $N = n_1 + n_2 - 2$.

Inicialmente trataremos da expansão assintótica das probabilidades de má classificação, não condicionadas às amostras retiradas das populações Π_1, Π_2.

De acordo com o capítulo 1, a regra de classificação usando W(x) é definida:

Classifica x em Π_1 se $W(x) > c$,e caso contrário classifica em Π_2 onde c é uma constante (depende dos custos de má classificação e das probabilidades a priori das populações Π_1, Π_2).

Temos então as probabilidades (não condicionadas) de má classificação:

$$P_1^* = P(W(x) < c/x \in \Pi_1)$$

e,

$$P_2^* = P(W(x) > c/x \in \Pi_2).$$

A distribuição exata de W(x) é muito complicada, mas o teorema 1.4 trata da sua distribuição assintótica onde W(x) é distribuida $N(1/2\,\Delta^2, \Delta^2)$ se x pertence a Π_1 e distribuída $N(-1/2\,\Delta^2,\Delta^2)$ se x pertence a Π_2. Δ^2 é a distância ao quadrado de Mahalanobis entre as duas populações Π_1, Π_2.

TEOREMA 2.1

Com $n_1 \to \infty$, $n_2 \to \infty$ e $n_1/n_2 \to$ a; limite positivo,

(1)
$$P\left(\frac{W(x) - 1/2\,\Delta^2}{\Delta} \leq u/\ \Pi_1\right) =$$

$$= G(u) - g(u)\left(\frac{1}{2n_1\Delta^2}[u^3 + (p-3)u - p\Delta] + \frac{1}{2n_2\Delta^2}[u^3 + 2\Delta u^2 + (p-3+\Delta^2)u + (p-2)\Delta] \right.$$

$$\left. + \frac{1}{4N}[4u^3 + 4\Delta u^2 + (6p-6+\Delta^2)u + 2(p-1)\Delta] \right) + O(N^{-2})$$

e $$P_2^* = P(W(x) \geq c / \Pi_2) = P\left(\frac{W(x) + 1/2\Delta^2}{\Delta} \geq \frac{c + \Delta/2}{\Delta} \right) =$$

$$= P\left(\frac{-(W + 1/2\Delta^2)}{\Delta} \leq u / \Pi_2 \right)$$

é obtida de (1) trocando n_1 por n_2. Para P_1^*, $u = (c - 1/2\Delta^2)/\Delta$ e para P_2^* e $u = -(c + 1/2\Delta^2)/\Delta$. $G(\cdot)$ e $g(\cdot)$ são respectivamente Fd. e fd. da $N(0,1)$.

Fazendo $c = 0$, $u = -\Delta/2$ e $n_1 = n_2$ temos o seguinte corolário:

COROLÁRIO 2.1

$$(2) \qquad P_1^* = Pr\left(W \leq 0 / \Pi_1, \lim_{N\to\infty} \frac{n_1}{n_2} = 1 \right) =$$

$$= G(-\Delta/2) + 1/N\, g(\Delta/2)[(p-1)/\Delta + p/4\Delta] + o(N^{-1}) =$$

$$= Pr\left(W \geq 0 / \Pi_2, \lim_{N\to\infty} \frac{n_1}{n_2} = 1 \right),$$

Notamos que o termo de correção é positivo, isto é, a probabilidade de má classificação é maior que a aproximação normal.

Geralmente Δ^2 é desconhecido. A distância ao quadrado amostral de Mahalonobis,

(3) $$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

é um estimador viciado de Δ^2. A esperança de D^2 é:

(4) $$E(D^2) = \frac{N}{N-p-1}\left[\Delta^2 + p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right].$$

O estimador não viciado de Δ^2 é:

(5) $$\frac{(N-p-1)}{N} D^2 - p\left(\frac{n_1 + n_2}{n_1 n_2}\right)$$

Anderson [1958], apresenta o seguinte resultado utilizando D^2.

TEOREMA 2.2

Se $n_1/n_2 \to$ a limite positivo com $N \to \infty$

$$\Pr\left(\frac{W(x) - 1/2D^2}{D} \leq u / \Pi_1\right) =$$

$$= G(u) - g(u)\left(\frac{1}{n_1} \left(u/2 - (p-1)/\Delta\right) + \frac{1}{N}\left[u^3/4 + (p - 3/4)u\right]\right) + O(N^{-2}).$$

$$P\left(\frac{-(W(x) + 1/2\ D^2)}{D} \leq u/\ \Pi_2\right) = G(u) - g(u)\left(\frac{1}{n_2}\left(u/2 - (p-1)/\Delta\right) + \right.$$

$$\left. \frac{1}{N}\left[u^3/4 + (p- 3/4)u\right]\right) + O(N^{-2}).$$

Para um caso especial, suponhamos que estamos interessados em escolher um ponto c para o qual a probabilidade de má classificação seja controlada. Seja $\alpha = Pr\ [W(x) < c/\ \Pi_1]$. Anderson [1958] deriva o seguinte teorema:

TEOREMA 2.3

Seja u_0 tal que $G(u_0) = \alpha$ e seja,

$$u = u_0 - \frac{1}{n_1}\left[(p-1)/D - 1/2\ u_0\right] + \frac{1}{N}\left[(p-3/4)u_0 + 1/4\ u_0^3\right],$$

com $n_1 \to \infty$, $n_2 \to \infty$ e $n_1/n_2 \to$ a limite positivo. Então,

$$Pr\left(\frac{W(x) - 1/2\ D^2}{D} \leq u/\ \Pi_1\right) = \alpha + O\ (N^{-2}).$$

Portanto, para um ponto $c = Du + 1/2\ D^2$ nós obtemos uma probabilidade desejada de má classificação fixada α.

Agora avaliaremos a probabilidade condicional de má classificação após obtido duas amostras de observações multivariadas do Π_1, Π_2 com tamanhos n_1, n_2 respectivamente.

Sejam $\bar{x}_1$ e $\bar{x}_2$ médias amostrais no espaço p-dimensional e S a matriz de variâncias e covariâncias amostral. Condicionada a $\bar{x}_1$, $\bar{x}_2$, S, W é normalmente distribuida com médias:

$$\mu^*_1: E(W/\ \Pi_1,\ \bar{x}_1,\ \bar{x}_2,\ S) = [\ \mu_1 - 1/2\ (\bar{x}_1 + \bar{x}_2)]'S^{-1}\ (\bar{x}_1 - \bar{x}_2)$$

$$\mu^*_2: E(W/\ \Pi_2,\ \bar{x}_1,\ \bar{x}_2,\ S) = [\ \mu_2 - 1/2\ (\bar{x}_1 + \bar{x}_2)]'S^{-1}\ (\bar{x}_1 - \bar{x}_2)$$

e matriz de variâncias e covariâncias:

$$v^* = V(W/\bar{x}_1,\ \bar{x}_2,\ S) = (\bar{x}_1 - \bar{x}_2)'S^{-1}\ \Sigma\ S^{-1}\ (\bar{x}1 - \bar{x}2)$$

Notemos que tomando os limites dessas expressões acima, quando n_1, $n_2 \to \infty$ obtemos:

$$\lim_{n_1, n_2 \to \infty} \mu^*_1 = 1/2\ \Delta^2 \quad e \quad \lim_{n_1, n_2 \to \infty} \mu^*_2 = -\ 1/2\ \Delta^2$$

$$\lim_{n_1, n_2 \to \infty} V^* = \Delta^2.$$

Sejam P_1 e P_2 probabilidades condicionais de má classificação assim obtidas:

(7) $$P_1 = P_1[2/1,\ c,\ \bar{x}_1,\ \bar{x}_2,\ S] =$$

$$= G\left(\frac{c - \mu^*_1}{\sqrt{(V^*)}}\right)$$

e,

(8) $$P_2 = P_2[1/2,\ c,\ \bar{x}_1,\ \bar{x}_2,\ S] =$$

$$= 1 - G\left(\frac{c - \mu^*_2}{\sqrt{(V^*)}}\right).$$

O cálculo para se obter (7) e (8) é como se segue:

$$P_1 = P[W(x) \leqslant c \ / \ \bar{x}_1,\ \bar{x}_2,\ S,\ \Pi_1] =$$

$$= P\left(\frac{[x - 1/2(\bar{x}_1 + \bar{x}_2)]'S^{-1}(\bar{x}_1 - \bar{x}_2)] - [\mu_1 - 1/2(\bar{x}_1 + \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2)}{[(\bar{x}_1 - \bar{x}_2)'S^{-1}\Sigma\, S^{-1}(\bar{x}_1 - \bar{x}_2)]^{1/2}} \leqslant\right.$$

$$\left. \leqslant \frac{c - [\mu_1 - 1/2(\bar{x}_1 - \bar{x}_2)]'S^{-1}(\bar{x}_1 - \bar{x}_2)}{[(\bar{x}_1 - \bar{x}_2)'S^{-1}\Sigma\, S^{-1}(\bar{x}_1 - \bar{x}_2)]^{1/2}}\right) =$$

$$= G\left(\frac{c - \mu^*_1}{\sqrt{[V^*]}}\right)$$

de forma semelhante obtemos P_2.

P^*_1, P^*_2 (probabilidades não condicionais) são as esperanças de P_1, P_2 (probabilidades condicionais) quando $\bar{x}_1$, $\bar{x}_2$ variam.

A próxima secção trataremos de métodos de estimação de P_1, P_2.

2.3 - ESTIMADORES DE P_1 e P_2

Lachenbruch e Mickey [1968], fazem uma comparação entre os métodos que aqui são apresentados através de ensaios de Monte Carlo.

Um estudo assintótico desses métodos é proposto por Sorum [1971]. Kshirsagar [1972], desenvolve toda metodologia dos métodos propostos por Lachenbruch e Mickey.

Aqui nós trataremos de apresentar esses métodos de acordo com os trabalhos citados acima.

Suponhamos como na secção anterior que temos duas amostras de tamanhos n_1 e n_2 de observações multivariadas das populações Π_1, Π_2, normais com vetores de médias μ_1, μ_2, e matriz de variâncias e covariâncias Σ ; sendo μ_1, μ_2, Σ desconhecidos.

O procedimento de classificação é baseado na função de Anderson:

$$W(x) = (x - 1/2(\bar{x}_1 + \bar{x}_2))' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

Seja x uma observação da população de Π_1 ou Π_2. S, $\bar{x}_1$, $\bar{x}_2$ como já conhecidos, são as quantidades amostrais; o ponto c crítico para a classificação da observação x será aqui tomado como $c = 0$. Então, P_1, a probabilidade de má classificação (de (7)) pode ser escrita:

$$(9) \qquad P_1 = P\,[W(x) < 0 \;/\; x \in \Pi_1,\ \bar{x}_1,\ \bar{x}_2,\ S] =$$

$$= G\left(\frac{[-\mu_1+1/2(\bar{x}_1+\bar{x}_2)]'S^{-1}(\bar{x}_1-\bar{x}_2)]}{[(\bar{x}_1-\bar{x}_2)'S^{-1}\Sigma\, S^{-1}(\bar{x}_1-\bar{x}_2)]^{1/2}}\right)$$

de mesma forma por (8) obtemos P_2, a probabilidade de má-classificação, assim escrita:

$$P_2 = P[W(x) > 0/x \in \Pi_2, \bar{x}_1, \bar{x}_2, S] =$$

$$= 1 - G\left(\frac{[-\mu_2+1/2(\bar{x}_1+\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)]}{[(\bar{x}_1-\bar{x}_2)'S^{-1}\Sigma\, S^{-1}(\bar{x}_1-\bar{x}_2)]^{1/2}}\right).$$

Podemos dividir os métodos para estimar P_1 e P_2 em duas classes:

1 - métodos empíricos;

2 - método usando as propriedades da distribuição normal.

Os primeiros só utilizam as amostras, não sendo necessário conhecer distribuições das populações origem para avaliar a função discriminante, enquanto que para o segundo é necessário a propriedade de normalidade das populações origem.

Na primeira classe Lachenbruch e Mickey [1968], apresentam três métodos, o método H,U,R, (sendo o último um dos métodos mais usados para a estimação de P_1 e P_2). No capítulo 1 foi feita uma pequena apresentação dos dois métodos que aqui serão especificados, H e U.

Os métodos são descritos abaixo:

1 - MÉTODO U

Se as amostras iniciais são suficientemente grandes, escolhemos dois subconjuntos de observações de cada grupo, um para construir a função discriminante e o restante de observações para estimar a probabilidade de erro. O número de erro de cada grupo será distribuido binomialmente com a probabilidade P_1 e P_2.

Após essas estimativas terem sido obtidas, construimos a função discriminante com o uso de toda a amostra.

Este procedimento apresenta vários problemas:

a - talvez não seja possível obter grandes amostras;

b - a função que é avaliada não é a mesma que é usada;

c -existe o problema relacionado com o tamanho da amostra para avaliar a função discriminante, se for grande uma boa estimativa do resultado da função discriminante pode ser obtida, mas a função discriminante será pobre; se for pequena, a função discriminante será superior, mas a estimativa do resultado será altamente variável. Pelo fato de dividir as amostras de cada grupo, informações importantes podem ser perdidas na construção da função discriminante.

Lachenbruch e Mickey [1968] não incluem este método em seus trabalhos experimentais.

2 - MÉTODO R (RESUBSTITUIÇÃO)

As amostras das populações já escritas anteriormente são usadas para construir W(x). O método consiste em usar as mesmas amostras para avaliar a função W(x).

O estimador de P_1 é obtido pela proporção de observações da amostra de Π_1 mal classificadas por W(x). Da mesma maneira obtemos o estimador de P_2.

Este procedimento dá uma estimativa muito otimista (subestima P_1 e P_2), isto é porque as mesmas amostras usadas para a construção da função, são usadas para avaliar seu desempenho.

Os métodos a seguir são incluidos na segunda classe, isto é, necessitam da propriedade de normalidade das populações, excluindo o método U.

3 - MÉTODO D E DS

No capítulo 1, calculamos os valores das probabilidades de má classificação P_1 e P_2, considerando os parâmetros populacionais conhecidos e o resultado é $G(-\Delta/2)$ para ambas probabilidades.

No caso em que os parâmetros são desconhecidos usamos as quantidades amostrais $\bar{x}_1$, $\bar{x}_2$, S para estimar μ_1, μ_2, Σ em Δ^2, obtendo assim $D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$.

Substituindo D em $G(-\Delta/2)$ obtemos $G(-D/2)$ que é o estimador de P_1 e P_2. Este método é denotado como método D.

D^2 é um estimador viciado de Δ^2, como visto em (4), superestima Δ^2 consequentemente $G(-D/2)$ subestima $G(-\Delta/2)$.

Como modificação deste método usamos o estimador dado em (5) não viciado de Δ^2:

$$(10) \qquad D^{*2} = \frac{(N - p - 1)}{N} D^2 - p\left(\frac{n_1 + n_2}{n_1 n_2}\right).$$

Infelizmente, quando n_1, n_2 são relativamente pequenos em relação a p e, D^2 é pequeno, D^{*2} é frequentemente negativo. Para que isto seja evitado, usamos a quantidade,

(11) $$DS^2 = \left(\frac{N - p - 1}{N}\right) D^2$$

para estimar Δ^2, ignorando o termo $p\,(n_1 + n_2)/n_1 n_2$ da expressão de D^{*2}.

Assim ambos P_1 e P_2 podem ser estimados por $\Phi(-DS/2)$. Este é o método DS.

4 - MÉTODOS O E OS

Estes métodos de estimação são baseados nos resultados de Okamoto [1963]. Ele apresenta expansões assintóticas de $W(x)$ e obtem as probabilidades (não condicionais) assintóticas em função de n_1, n_2, p, Δ^2.

Ele apresenta o seguinte resultado para expansão assintótica de P_1^*:

(12) $$P_1^* = Pr[W(x) < 0 / x \in \Pi_1] =$$

$$= G\ (-1/2\ \Delta)\ + \frac{a_1}{n_1} + \frac{a_2}{n_2} + \frac{a_3}{N} + \frac{b_{11}}{n_1^2} + \frac{b_{22}}{n_2^2} +$$

$$+ \frac{b_{12}}{n_1 n_2} + \frac{b_{13}}{n_1 N} + \frac{b_{23}}{n_2 N} + \frac{b_{33}}{N^2} + O_3$$

Obtemos $P_2^* = P[W(X) > 0 / x \in \Pi_2]$ pela troca de n_1 e n_2 na expressão (12). Os valores dos a's e b's são em função dos parâmetros de Π_1 e Π_2; ele dá também o valor tabulado dos a's e b's para um caso especial.

Podemos obter o estimador de P_1, substituindo na expressão (12), D em Δ. De mesma forma, obtemos o estimador de P_2.

Este procedimento é denotado como método Q. Outra forma de obter através da expressão (12) o estimador de P_1 é usar o estimador DS (11) no lugar de Δ, De mesma forma, estima P_2. Este é denominado método OS.

5 - MÉTODO U

Faremos o uso de métodos empíricos mas não teremos sérios vícios como os métodos R e H. De todo conjunto amostral (n_1+n_2) das populações Π_1 e Π_2, omitimos uma observação das amostras de Π_1 ou Π_2. Construimos a função discriminante baseada nas $n_1+ n_2- 1$ observações.

Suponhamos que omitimos um vetor $x_j (j=1,\ldots,n_k)$ da amostra de Π_k, k=1,2.

Sejam $U_j = x_j - x_k$ e $c_k = \dfrac{n_k}{(n_k - 1)N}$

onde $N = n_1 + n_2 - 2$

e, $a_j = U_j' S^{-1} U_j$.

A função de classificação de x_j construída sem esta observação é dada por:

$$(13) \quad D_j^*(x_j) = \frac{n_1 + n_2 - 3}{n_1 + n_2 - 2} \left(x_j - 1/2(\bar{x}_1 + \bar{x}_2) + U_j / 2(n_k - 1) \right)' S_j^{-1} \left((\bar{x}_1 - \bar{x}_2) + (-1)^k U_j/(n_k - 1) \right)$$

onde $\bar{x}_1$, $\bar{x}_2$ e S são obtidas pelas amostras inteiras de Π_1 e Π_2, e $S_j^{-1} = S + (c_k S^{-1} U_j U_j' S^{-1})/(1 - c_k a_j)$.

Neste método seria necessário em cada retirada de uma observação a construção da matriz inversa S^{-1}. Seria muito dispendiosa a parte computacional. O método foi desenvolvido da forma que requer somente uma inversão da matriz de variâncias e covariâncias amostrais.

Este procedimento de inversão foi baseado nos estudos de Bartlett [1952], da seguinte forma:

se $B = A + uv'$ então $B^{-1} = A^{-1} - (A^{-1} uv' A^{-1} / 1 + v' A^{-1} u)$

onde B e A são matrizes quadradas não singulares e u e v são vetores colunas. Por este raciocínio é obtida S_j^{-1}.

Para obter o estimador de P_1, omitimos uma observação da amostra de Π_1 e fazemos k = 1 em (13).

De acordo com as regras de classificação x_j será corretamente classificado em Π_1 se

$$D_j^*(x_j) \geqslant 0$$

e será classificado em Π_2 se,

$$D_j^*(x_j) < 0.$$

Sucessivamente seguimos este procedimento para todo o conjunto de observações da amostra de Π_1.

A proporção m_1/n_1 de m_1 observações mal classificadas é o estimador de P_1.

Para estimar P_2, o procedimento é similar, omitimos sucessivamente uma observação da amostra de Π_2 e seguimos o mesmo raciocínio para a construção da função discriminante dada em (13), fazendo k = 2. Obtemos então, o estimador m_2/n_2 de P_2, que é proporção de observações na amostra de Π_2 mal classificadas.

6 - MÉTODO $\bar{U}$

Lachenbruch e Mickey [1968] propõem, baseado no método U, $G(-\bar{D}_1/s(D_1))$ como estimador de P_1 e $G(\bar{D}_2/s(D_2))$ como estimador de P_2 onde,

$$\bar{D}_1 = \frac{1}{n_1} \sum_{r=1}^{n_1} D_r^*(x_{1r})$$

$$\bar{D}_{(2)} = \frac{1}{n_2} \sum_{r=1}^{n_2} D_r^*(x_r).$$

$s^2_{(D_1)}$ é variância amostral dos n_1's valores de $D_r^*(x_r)$.

$s^2_{(D_2)}$ é variância amostral dos n_2's valores de $D_r^*(x_r)$.

Este método combina com as características de um método empírico e faz uso da distribuição normal.

RESULTADOS EXPERIMENTAIS

Das oito técnicas de estimação de P_1 e P_2, Lachenbruch e Mickey [1968], fazem experimentos amostrais através de ensaios de Monte Carlo envolvendo sete delas, não considerando o método H. Eles tomam um total de 288 amostras de populações normais com $\Sigma = I$ e $\mu_1 = 0$; $\mu_2 = (\delta, 0, 0, \ldots, 0)$, sendo assim possível determinar a probabilidade exata de má classificação de cada amostra.

Essas amostras foram agrupadas seguindo tres critérios diferentes a serem observados nas tabelas 1, 2 e 3. Na tabela 1, as amostras são agrupadas com respeito à distância δ entre as duas populações. Para cada um dois seis valores de δ foram geradas 48 amostras. Na tabela 2, as amostras se distribuem com respeito à combinação de n_1, n_2 e p sendo, n_1 e n_2 o número de obsevações das populações 1 e 2 respectivamente e p a dimensão do espaço. Para cada uma das combinações n_1 e p estudadas, temos n_2 assumindo os valores n_1, $2n_2$, $3n_2$.

Em cada tripla (n_1, n_2, p) foram geradas 12 amostras sendo duas de cada um dos valores de δ . Como exemplo, tomando p=2 e $n_1 = n_2 = 4$, seis amostras digamos, A_1, $A_2 \ldots A_6$ foram geradas da população 2 de tamanho $n_2 = 4$ considerando que, para cada A_i uma distância $\delta/(i=1\ldots6)$. A mesma idéia para as amostras de tamanho $n_1 = 4$ da população 1. Portanto, obtemos um total de 12 amostras em $n_1 = n_2 = 4$ e p=2. A tabela 3, apresenta o número de amostras distribuídas segundo diversos valores do erro (e), p e m onde $m = n_1 + n_2$. O erro é classificado em 5 intervalos (e<0.05, 0.05<e<0.1, ... , e>0.20).

O conjunto de 288 amostras está dividido em 8 classes sendo que, cada classe é definida para valores de m-p-1. Em cada valor de p(2,4,8) as classes são definidas por m-p-1≤20, m-p-1>20 e para p=20 as classes são m-p-1≤25 e m-p-1>25. Segue um exemplo, para verificar como estas amostras estão divididas de acordo com o valor de m-p-1. Para p=2 e m-p-1≤20, verificando na tabela 2 que os pares s (n1=4, n2=4), (n1=4, n2=8), (n1=4, n2=12) satisfazem a condição m-p-1<20, temos então um total de 36 amostras nos tres pares.

Para finalizar, a tabela 3 apresenta os resultados experimentais dos números de amostras de cada método (O, OS, D, DS, U, Ū, R) dentro das classes dos erros (e). No exemplo anterior, para o método O, verificamos que do total de 36 amostras 14 estão em (e<0.05), 4 (0.05<e<0.1), 4 (0.1<e<0.15), 5 (0.15<e<0.2) e 9 (e>0.20).

TABELA 1

NÚMERO DE AMOSTRAS PARA UMA DISTÂNCIA δ^2

δ^2	1.098	1.817	2.836	4.293	6.574	11.482
$G(-\delta/2)$	.30	.25	.20	.15	.10	.05
n	48	48	48	48	48	48

TABELA 2

TAMANHOS AMOSTRAIS COM NÚMERO DE PARÂMETROS

	p			
	2	4	8	20
	4.4	8.8	8.8	15.15
	4.8	8.16	8.16	15.30
n_1, n_2	4.12	8.24	8.24	15.45
	16.16	20.20	20.20	25.25
	16.32	20.40	20.40	25.50
	16.48	20.60	20.60	25.75

TABELA 3

NÚMERO DE AMOSTRAS DENTRO DE CADA ERRO

$$m = n_1 + n_2$$

	e<0.05	.05<e<.10	.10<e<.15	.15<e<.20	e>.20		T.AMOSTRAIS
O	97	68	42	28	53	Todos	288
OS	118	81	45	23	21		
D	87	52	37	40	72		
DS	94	73	50	36	35		
Ü	107	70	48	25	38		
U	73	82	62	29	42		
R	64	60	46	30	88		
O	14	4	4	5	9	p = 2	36
OS	8	10	6	6	6		
D	14	3	5	5	9	m - p - 1≤20	
DS	11	8	6	4	7		
Ü	11	8	6	0	11		
U	7	6	6	3	14		
R	6	8	5	3	14		
O	10	7	3	2	2	p = 4	24
OS	7	11	2	2	2		
D	11	5	4	2	2	m - p - 1≤20	
DS	11	6	4	1	2		
Ü	10	3	5	3	3		
U	5	8	6	1	4		
R	6	4	8	1	5		

	e<0.05	.05<e<.10	.10<e<.15	.15<e<.20	e>.20		T.AMOSTRAIS
O	3	7	3	5	6	p = 8	24
OS	9	5	4	1	5		
D	1	4	7	5	7	m - p - 1≤20	
DS	6	6	2	3	7		
Ū	7	3	3	2	9		
U	5	4	3	6	6		
R	2	2	5	6	9		
O	4	4	2	4	10	p = 20	24
OS	12	4	5	1	2		
D	1	0	3	7	13	m - p - 1≤25	
DS	1	4	8	3	8		
Ū	9	6	6	0	3		
U	6	6	5	4	3		
R	1	1	3	3	16		
O	15	11	8	1	1	p = 2	36
OS	16	11	6	2	1		
D	17	8	8	2	1	m - p - 1>20	
DS	17	8	8	2	1		
Ū	19	8	5	3	1		
U	14	9	10	2	1		
R	16	8	7	4	1		

	e<0.05	.05<e<.10	.10<e<.15	.15<e<.20	e>.20		T.AMOSTRAIS
O	21	15	9	1	2	p = 4	48
OS	26	9	9	2	2		
D	22	15	3	6	2	m - p - 1>20	
DS	23	14	7	3	1		
Ü	24	12	5	5	2		
U	15	15	10	5	3		
R	17	16	8	2	5		
O	22	9	9	4	4	p = 8	48
OS	23	15	5	5	0		
D	17	12	5	4	10	m - p - 1>20	
DS	18	19	3	7	1		
Ü	16	19	6	4	3		
U	9	18	15	3	3		
R	13	12	4	7	12		
O	8	11	4	6	19	p = 20	48
OS	17	16	8	4	3		
D	4	5	2	9	28	m - p - 1>25	
DS	7	8	12	13	8		
Ü	11	11	12	8	6		
U	12	16	7	5	8		
R	3	9	6	4	26		

Na tabela 3, $e = |P_i - P_{iK}|$, onde P_i é verdadeira probabilidade de má classificação e P_{iK} é o estimador de P_i, para $K = O, OS, U, US, DS, \bar{U}, U, R$.

A figura 1, é o gráfico dos métodos em relação aos valores acumulados das amostras com as classes dos e's, sendo estes truncados em 0,20.

As conclusões dos autores são que os métodos mais usados são significativamente mais pobres que os novos métodos por eles propostos.

Alguns pontos aqui são colocados por eles observando os resultados do experimento.

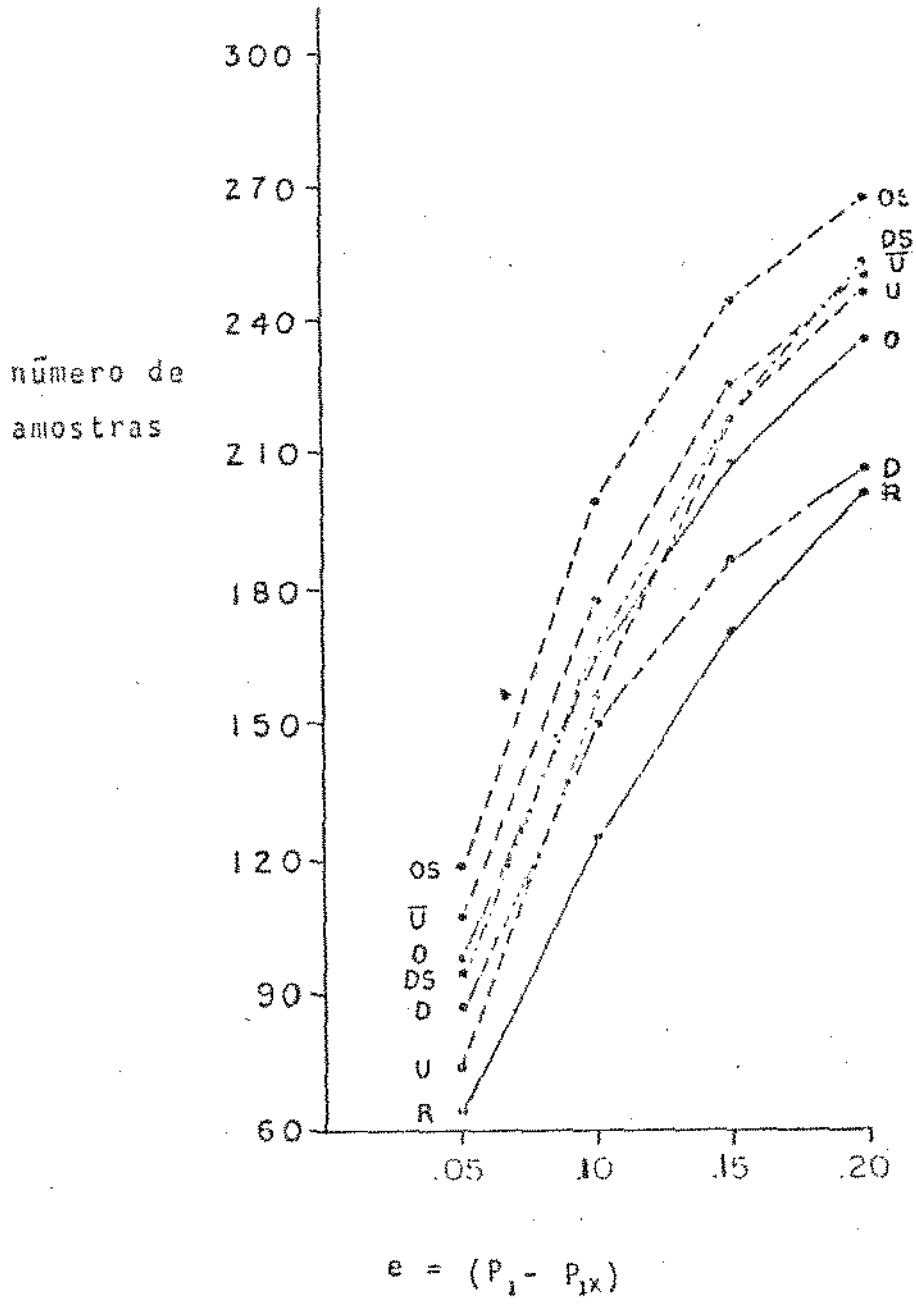
Para p = 20, o método OS, $\bar{U}$ e U foram superiores aos outros. Para o caso de pequenas dimensões, os métodos OS, $\bar{U}$ e U foram aproximadamente equivalentes. Para p = 2, 4 e 8, os métodos D e O tiveram resultados similares, no caso de p = 20, o método O foi algumas vezes superior.

Os métodos baseados na distribuição normal OS, $\bar{U}$ e DS são superiores aos outros.

Se D^2 ou n_1 e n_2 são pequenos o método OS não é recomendado devido à dificuldade com a aproximação de Okamoto. Neste caso, o método $\bar{U}$ pode ser usado se a aproximação da normalidade for assumida, caso contrário o método U pode ser satisfatório. Para grandes valores de p, os métodos OS, $\bar{U}$ ou U podem ser usados. Suas conclusões são que os métodos R e D são relativamente pobres.

Para grandes valores de p, o método O é substancialmente mais pobre que os métodos OS, $\bar{U}$ e U. O método OS parece ser geralmente preferido ao método O e da mesma forma, o método DS sobre o método D. Se a aproximação da normalidade pode ser assumida, os métodos OS e $\bar{U}$ são bons.

Os autores também apresentam resultados experimentais para os métodos R e U em que a razão do tamanho amostral não parece afetar os resultados destes métodos empíricos.



Cohran [1968], faz um rápido comentário sobre o trabalho de Lachenbruch e Mickey [1968] que aqui foi apresentado. Ele comparou os resultados da tabela 3 assumindo escores 1, 3, 5, 7, 10 nas classes (e<0.05, 0.05<e<0.10, ... , e>0.2 respectivamente calculando assim os valores médios. Cmo exemplo, para o método O e p=20, m-p-1≤25, a média é igual a $\sum_i f_i x_i / \sum_i f_i$ onde x_i=(1, 3, 5, 7, 10) para os respectivos valores de f_i(4, 4, 2, 4, 10) i=1, ... ,5.

Nas quatro primeiras colunas, os métodos são colocados em ordem com respeito aos valores médios dos escores, para p = 20 e p = 8.

TABELA 4

p = 20 m - p - 1≤25	p = 20 m - p - 1>25	p = 8 m - p - 1≤20	p = 8 m - p - 1>20	MÉDIAS p=20,8	MÉDIAS p=4,2
OS 3,2	OS 3,4	OS 4,2	OS 2,7	OS 3,4	DS 3,3
$\bar{U}$ 3,6	U 4,4	DS 5,2	DS 3,1	$\bar{U}$ 4,3	OS 3,5
U 4,5	$\bar{U}$ 4,6	U 5,6	$\bar{U}$ 3,4	U 4,6	O 3,5
DS 6,4	DS 5,5	O 5,6	O 3,4	DS 5,0	D 3,5
O 6,4	O 6,1	$\bar{U}$ 5,6	U 3,9	O 5,4	$\bar{U}$ 3,6
D 8,1	R 7,2	D 6,4	D 4,3	D 6,6	U 4,3
R 8,3	D 7,7	R 6,9	R 5,0	R 6,8	R 4,4

O método OS é superior em todos os casos e não existe muita diferença entre os métodos U e $\bar{U}$. Para p = 20 o método DS é inferior aos métodos OS, $\bar{U}$ e U sendo que para p = 8 ele fica na segunda posição.

A segunda parte da tabela mostra separadamente os pontos médios para p = 20,8 e p = 2,4. Os resultados mostram que para pequenos valores de p(2,4) as diferenças não são muito significativas entre os métodos. Cochran concorda com os autores sobre a superioridade dos métodos OS e DS.

Outro trabalho envolvendo estes estimadores é proposto por Sorum [1971]. Ele considera o problema de estimação da probabilidade de má classificação P_2 em dois grupos de populações normais multivariadas com vetores de médias desconhecidas e matriz de covariância conhecida, utilizando a função de classificação de Anderson com base nas amostras. Além dos estimadores propostos por Lachenbruch e Mickey (O, OS, $\bar{U}$, U, D, DS, R), ele apresenta estimadores baseado em amostras testes. Alguns estimadores são desenvolvidos considerando populações normais univariadas. Seus estudos são baseados no Erro Quadrático Médio Assintótico (condicional e não condicional).

Em suas conclusões, o uso da normalidade não proporciona um melhor estimador, pois o método D é pobre e em particular é mais pobre que o método U. Seus resultados também confirmam que o método R é o mais pobre entre todos apresentados, mas mostra que os métodos D, DS, O e OS são assintóticamente equivalentes.

CAPÍTULO 3

COMPARAÇÃO DE MÉTODOS NA ESTIMAÇÃO DO VÍCIO DO ESTIMADOR DA RAZÃO DE ERRO REAL

3.1 - INTRODUÇÃO

A Razão de Erro Real associada a uma função discriminante é dada por P_1 e P_2, as probabilidades de má classificação vistas no capítulo 1 e 2.

Para estimar a Razão de Erro Real, nós usamos o método R já estudado no capítulo 2, conhecido em outras bibliografias e aqui também denotado pela Razão de Erro Aparente.

Estudos mostram que este método subestima a Razão de Erro Real. O nosso objetivo é estudar seu vício, no contexto de comparar dois métodos de estimação deste vício com o estimador bootstrap, proposto por Efron B. [1979].

Os outros dois estimadores aqui estudados são o cross-validation e o paramétrico. O primeiro também é apresentado por Efron [1979] e o segundo é estudado por Mclachlan G. J. [1976].

Um estudo de comparação dos três métodos acima para estimar o vício da Razão Erro Aparente é apresentado em outro trabalho de Mclachlan [1980].

O capítulo está dividido em secções onde a secção 3.2 apresenta a metodologia do nosso trabalho. A descrição dos métodos de estimação do vício é vista na secçao 3.3, os resultados experimentais com os três métodos junto com a metodologia para a obtenção desses resultados, são apresentados na secção 3.4. A conclusão dos resultados experimentais é vista no capítulo 4.

3.2 - METODOLOGIA DOS ESTUDOS DO VÍCIO DA RAZÃO ERRO APARENTE

Sejam F_1 e F_2 as distribuições das populações Π_1, Π_2 e H_1 e H_2 amostras de tamanhos n_1 e n_2 das populações Π_1 e Π_2 respectivamente. Com base nas amostras H_1 e H_2 é construída a função discriminante de Anderson $W(x)$.

As distribuições F_1 e F_2 são pertencentes ao espaço p-dimensional com vetores de médias μ_1 e μ_2, e matrizes de variâncias e covariâcias iguais a Σ, sendo Σ , μ_1 e μ_2 quantidades desconhecidas. Como já visto nos capítulos anteriores classifica- se uma nova obsevação x em Π_1 se $W(x) > c$, caso contrário, classifica em Π_2. Aqui tomaremos $c = 0$ e $F_i \sim N(\mu_i, \Sigma)$, $i = 1,2$.

A Razão de Erro Real é dada por:

$$P_i = \Phi\left(\frac{(-1)^i \left[\mu_i - 1/2(\bar{x}_1 + \bar{x}_2)\right]' S^{-1} (\bar{x}_1 - \bar{x}_2)}{\left[(\bar{x}_1 - \bar{x}_2)' S^{-1} \Sigma S^{-1} (\bar{x}_1 - \bar{x}_2)\right]^{1/2}}\right)$$

onde P_i é a probabilidade de má classificação e $i = 1,2$.

Os estudos aqui são desenvolvidos para o caso em que $i = 2$, isto é, no cálculo de P_2, probabilidade de má classificação de uma obsevação de Π_2 em Π_1. R_2 é o estimador Razão Erro Aparente de P_2, pelo método R, que é a proporção de observações em H_2 mal classificadas por W. Já que R_2 subestima P_2, estudaremos o vício B_2, isto é,

$$B_2 = E(P_2 - R_2)$$

A próxima secção, apresenta os estimadores de B_2 pelos métodos bootstrap, cross-validation e paramétrico.

3.3 - ESTIMADORES DE B_2

Os três métodos aqui aparesentados são divididos em duas classes. Na primeira classe estão os métodos empíricos,bootstrap e cross-validation, os quais não necessitam do conhecimento das distribuições de F_1 e F_2. Na segunda classe, está o paramétrico, é necessária a suposição de normalidade de F_i, para a validade do método.

1 - MÉTODO BOOTSTRAP

1º passo:

Duas amostras com reposição H_1^* e H_2^* são geradas das populações Π_1 e Π_2, baseadas nas amostras originais H_1 e H_2, colocando massa pontual $1/n_i$ em cada ponto x_{ij} $(j = 1,\ldots,n_i)$ em H_i $(i = 1,2)$ (amostras bootstrap).

2º passo:

Baseado nas novas amostras H_1^* e H_2^*, construímos a função discriminante $W^*(x)$ substituindo $\bar{x}_1$, $\bar{x}_2$ e S em $W(x)$ por $\bar{x}_1^*$, $\bar{x}_2^*$ e S^*

3º passo:

Avaliamos cada observação x_{2j} $(j = 1,\ldots,n_2)$ de H_2^* na função de $W^*(x)$.

4º passo:

A Razão de Erro Aparente de $W^*(x)$, R_2^* é calculada como

$$R_2^* = \frac{m_2^*}{n_2}$$

onde m_2^* é o número de observações em H_2 mal classificados por $W^*(x)$

5º passo:

Calculamos a diferença $d = R_2^{*-} R_2^*$, onde R_2^* é a proporção de membros em H_2 mal classificados por $W^*(x)$.

6º passo:

Repetimos os procedimentos anteriores em número n (grande) de vezes, até obtermos d_1, d_2,..., d_n.

7º passo:

O estimador de B_2 pelo método acima é calculado por:

$$b_2 = \sum_{i=1}^{n} \frac{d_i}{n}$$

2 - MÉTODO CROSS-VALIDATION (c_2)

O procedimento a seguir é equivalente ao já apresentado no capítulo 2, método U .

1º passo:

Eliminamos uma observação de $H_2(x)$. Suponhamos x_{2j} para $j=1,2...n_2$.

2º passo:

Retirado x_{2j} de H_2, construímos a função discriminante $W_j(x)$ de mesma forma que foi construída $W(x)$, só que, agora considerando n_2-1 observações de H_2 e n_1 observações de H_1.

3º passo:

Avaliamos a observação x_{2j} em $W_j(x)$, verificando se x_{2j} é ou não mal classificada.

Após repetirmos estes procedimentos com todas as observações de H_2, calculamos a proporção m_{2c}/n_2 de observações de H_2 mal classifcadas pela função W.

O estimador de B_2 obtido através deste método é:

$$c_2 = \frac{m_{2c}}{n_2} - R_2$$

onde R2 é o estimador Razão de Erro Aparente associado a H_1, H_2 e W(x).

3 - ESTIMADOR PARAMETRICO

Mclachlan [1976], apresenta um estimador do vício assintótico da Razão de Erro Aparente. Em seus estudos, $F_i \sim N(\mu_i, \Sigma)$; i = 1,2; ambos μ_i e Σ desconhecidos.

A regra de classificação por ele usada é a mesma por nós estudada, W(x), a função de classificação de Anderson.

Segue o seguinte resultado por ele desenvolvido:

TEOREMA 3.1

O vício assintótico de R_2 é dado por:

$$\text{Vício}(R_2) = g(-1/2\,\Delta)[(1/4\,\Delta + (p-1)/\Delta)/n_2 + 1/2\,(p-1)\Delta/N]$$

até o termo de segunda ordem, com respeito $(n_1^{-1}, n_2^{-1}, N^{-1})$; onde

$$\Delta^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2),$$

$N = n_1 + n_2 - 2$ e g função de densidade da N(0,1).

De acordo com as amostras de H_1 e H_2 substituímos Δ^2 por $D^2 = (\bar{x}_1 - \bar{x}_2)' \bar{S}^{-1} (\bar{x}_1 - \bar{x}_2)$ obtendo assim,

$$m_2 = g(-1/2\ D)\left((1/4\ D + (p-1)/D\ )/n_2 + 1/2(p-1)D/N \right)$$

que é o estimador não viciado do vício (R_2) até o termo de segunda ordem .

3.4 - RESULTADOS DOS TRABALHOS EXPERIMENTAIS

Amostras das duas populações Π_1 e Π_2 com distribuições $F_i \sim N(\mu_i, \Sigma)$ para vários valores de p foram geradas por ensaios de Monte-Carlo. Valores diferentes para os vetores de médias μ_i; $i = 1,2$; foram assumidos e a matriz de variâncias e covariâncias Σ é a identidade.

A função discriminante usada é a de Anderson,

$$W(x) = [(x - 1/2\ (\bar{x}_1 + \bar{x}_2))]' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

com o ponto crítico c = 0.

A Razão de Erro Real P_2, assim como R_2, b_2, c_2 e m_2 são obtidos nas condições acima.

Para diferentes valores de n_1 e n_2= 10, 20, 40 e p = 2, 4, 6, 8; distintos valores de Δ (distância de Mahalanobis entre as duas populações) foram combinados.

Os valores de Δ com respectivos valores de μ_i; $i = 1,2$; n_1, n_2 e p estão apresentados nas tabelas.

A tabela 1 apresenta os valores médios da diferença entre A Razão de Erro Real e a Razão de Erro Aparente, $P_2 - R_2$; e dos estimadores bootstrap (b_2), cross-validation (c_2) e o paramétrico (m_2), com seus respectivos desvios médios. Os valores médios dos métodos assim como seus desvios foram obtidos através dos seguintes métodos iterativos:

média: $\bar{x}(i) = [(i-1)\bar{x}(i-1) + x(i)]/i$

desvio: $S = [s(n)/(n-1)]^{1/2}$ onde $s(1) = s(i-1) + [(i-1)(x(i)- \bar{x}(i-1))^2]/i$

para $x(i)$, $i=1,...n$.

Uma série de 100 independentes ensaios de Monte Carlo foram realizados para obter esses valores, sendo que, para os valores médios de b_2, 1000 replicações independentes do Monte Carlo foram realizadas em cada um dos 100 ensaios de Monte-Carlo. O processo de Monte Carlo foi realizado segundo a rotina apresentada em Goelzer [1970], escrita em Fortran4 - Vax11/785 para a distribuição normal (0,1). Sub-rotinas foram criadas para os métodos bootstrap, cross-validation e paramétrico assim como, para o cálculo da inversa e da probabilidade da normal reduzida. O programa da inversa é baseado no método decomposição de Cholesky. O método da inversa e o programa para o cálculo da probabilidade normal reduzida são apresentados em Kennedy & Gentle [1980].

É feito o cálculo da eficiência relativa dos desvios médios entre os métodos bootstrap e paramétrico, assim definido:

$$\varepsilon = \frac{s\ D\ (m_2)}{s\ D\ (b_2)}$$

Para finalizar, a tabela 2 apresenta os valores médios do quadrado da diferença $P_2 - T$, onde T assume os valores dos estimadores de P_2 incluíndo o vício, isto é, $T = R_2$, $R_2 + b_2$, $R_2 + m_2$, $R_2 + c_2$; $(P_2-T)^2$ é o estimador do erro quadrático médio de T

TABELA 1

As sequências dos resultados é baseada na dimensão do espaço p.

1.1) p = 2

1.1.1) Δ = 1 $\mu_1 = (0.5, 0.0)$

$\mu_2 = (-0.5, 0.0)$

	$n_1 = n_2 = 10$			$n_1 = n_2 = 20$		
	valores médios	desvios		valores médios	desvios	
$\hat{B}_2$	0,047	0,148		0,023	0,114	
b_2	0,057	0,022	$\varepsilon_1 = 1,2$	0,029	0,012	$\varepsilon_2 = 0,75$
m_2	0,057	0,026		0,028	0,009	
c_2	0,070	0,078		0,033	0,039	

1.1.2) Δ = 2 $\mu_1 = (1.0, 0.0)$

$\mu_2 = (-1.0, 0.0)$

	$n_1 = n_2 = 10$			$n_1 = n_2 = 20$		
	valores médios	desvios		valores médios	desvios	
$\hat{B}_2$	0,045	0,107		0,018	0,083	
b_2	0,036	0,017	$\varepsilon_3 = 0,44$	0,017	0,0078	$\varepsilon_4 = 0,34$
m_2	0,036	0,0074		0,018	0,0026	
c_2	0,049	0,076		0,017	0,029	

1.2) p = 4

1.2.1) Δ = 1 $\mu_1 = (0.0, 0.0, 0.5, 0.5)$

$\mu_2 = (-0.5, -0.5, 0.0, 0.0)$

	$n_1 = n_2 = 20$			$n_1 = n_2 = 40$		
	valores médios	desvios		valores médios	desvios	
$\hat{B}_2$	0,061	0,11		0,049	0,085	
b_2	0,062	0,014	$\varepsilon_5 = 1,1$	0,032	0,0077	$\varepsilon_6 = 0,7$
m_2	0,066	0,015		0,033	0,0054	
c_2	0,064	0,052		0,032	0,027	

1.2.2) $\Delta = 2$ $\mu_1 = (0.0, 0.0, 1.0, 1.0)$

$\mu_2 = (-1.0, -1.0, 0.0, 0.0)$

	$n_1 = n_2 = 20$			$n_1 = n_2 = 40$		
	valores médios	desvios		valores médios	desvios	
B_2	0,040	0,08		0,022	0,057	
b_2	0,038	0,012	$\varepsilon_7 \cong 0,6$	0,020	0,0058	$\varepsilon_8 \cong 0,4$
m_2	0,041	0,007		0,021	0,0025	
c_2	0,043	0,043		0,021	0,024	

1.3) p = 6

1.3.1) $\Delta = 1,2$ $\mu_1 = (0.0, 0.0, 0.0, 0.5, 0.5, 0.5)$

$\mu_2 = (-0.5, -0.5, -0.5, 0.0, 0.0, 0.0)$

	$n_1 = n_2 = 20$			$n_1 = n_2 = 40$		
	valores médios	desvios		valores médios	desvios	
B_2	0,081	0,093		0,053	0,069	
b_2	0,082	0,016	$\varepsilon_9 = 1,13$	0,044	0,0088	$\varepsilon_{10} = 0,76$
m_2	0,088	0,018		0,046	0,0067	
c_2	0,105	0,063		0,054	0,033	

1.3.2) $\Delta \cong 2,5$ $\mu_1 = (1.0,1.0,1.0,2.0,2.0,2.0)$

$\mu_2 = (2.0,2.0,2.0,1.0,1.0,1.0)$

	$n_1 = n_2 = 20$			$n_1 = n_2 = 40$		
	valores médios	desvios		valores médios	desvios	
$\hat{B}_2$	0,054	0,068		0,029	0,049	
b_2	0,048	0,015	$\varepsilon_{11} = 0,8$	0,024	0,008	$\varepsilon_{12} \cong 0,5$
m_2	0,049	0,012		0,025	0,0038	
c_2	0,062	0,060		0,027	0,024	

1.4) p = 8

1.4.1) $\Delta \cong 1,4$ $\mu_1 = (0,0,0,0,0.5,0.5,0.5,0.5)$

$\mu_2 = (-0.5,-0.5,-0.5,-0.5,0,0,0,0)$

	$n_1 = n_2 = 20$			$n_1 = n_2 = 40$		
	valores médios	desvios		valores médios	desvios	
$\hat{B}_2$	0,109	0,103		0,064	0,075	
b_2	0,094	0,017	$\varepsilon_{13} = 1,17$	0,054	0,0095	$\varepsilon_{14} \cong 0,8$
m_2	0,102	0,02		0,056	0,0078	
c_2	0,123	0,071		0,062	0,036	

1.4.2) $\Delta \cong 2,8$ $\mu_1 = (1.0,1.0,1.0,1.0,2.0,2.0,2.0,2.0)$

$\mu_2 = (2.0,2.0,2.0,2.0,1.0,1.0,1.0,1.0)$

	$n_1 = n_2 = 20$			$n_1 = n_2 = 40$		
	valores médios	desvios		valores médios	desvios	
$\hat{B}_2$	0,065	0,060		0,031	0,041	
b_2	0,049	0,018	$\varepsilon_{15} = 0,95$	0,026	0,0084	$\varepsilon_{16} = 0,64$
m_2	0,046	0,017		0,027	0,0054	
c_2	0,072	0,053		0,029	0,025	

TABELA 2

Resultados dos Erros Quadráticos Médios.

2.1) p = 2

2.1.1) Δ = 1

	$n_1 = n_2 = 10$	$n_1 = n_2 = 20$
$\widehat{EQM}$ (R_2)	0,024	0,013
$\widehat{EQM}$ ($R_2 + b_2$)	0,024	0,014
$\widehat{EQM}$ ($R_2 + m_2$)	0,024	0,014
$\widehat{EQM}$ ($R_2 + c_2$)	0,026	0,014

2.1.2) Δ = 2

	$n_1 = n_2 = 10$	$n_1 = n_2 = 20$
$\widehat{EQM}$ (R_2)	0,013	0,0072
$\widehat{EQM}$ ($R_2 + b_2$)	0,013	0,0073
$\widehat{EQM}$ ($R_2 + m_2$)	0,012	0,0071
$\widehat{EQM}$ ($R_2 + c_2$)	0,015	0,0070

2.2) p = 4

2.2.1) Δ = 1

	$n_1 = n_2 = 20$	$n_1 = n_2 = 40$
$\widehat{EQM}$ (R_2)	0,016	0,0095
$\widehat{EQM}$ ($R_2 + b_2$)	0,013	0,0080
$\widehat{EQM}$ ($R_2 + m_2$)	0,014	0,0080
$\widehat{EQM}$ ($R_2 + c_2$)	0,012	0,0081

2.2.2) Δ = 2

	$n_1 = n_2 = 20$	$n_1 = n_2 = 40$
$\widehat{EQM}$ (R_2)	0,0078	0,0037
$\widehat{EQM}$ ($R_2 + b_2$)	0,0068	0,0034
$\widehat{EQM}$ ($R_2 + m_2$)	0,0068	0,0034
$\widehat{EQM}$ ($R_2 + c_2$)	0,0067	0,0039

2.3) p = 6

2.3.1) $\Delta \cong 1,2$

	$n_1 = n_2 = 20$	$n_1 = n_2 = 40$
$\widehat{EQM}$ (R_2)	0,015	0,0076
$\widehat{EQM}$ $(R_2 + b_2)$	0,010	0,0053
$\widehat{EQM}$ $(R_2 + m_2)$	0,011	0,0053
$\widehat{EQM}$ $(R_2 + c_2)$	0,012	0,0047

2.3.2) $\Delta \cong 2,5$

	$n_1 = n_2 = 20$	$n_1 = n_2 = 40$
$\widehat{EQM}$ (R_2)	0,0075	0,0032
$\widehat{EQM}$ $(R_2 + b_2)$	0,0055	0,0028
$\widehat{EQM}$ $(R_2 + m_2)$	0,0056	0,0025
$\widehat{EQM}$ $(R_2 + c_2)$	0,0064	0,0030

2.4) p = 8

2.4.1) $\Delta \cong 1,4$

	$n_1 = n_2 = 20$	$n_1 = n_2 = 40$
$\widehat{EQM}$ (R^2)	0,022	0,0097
$\widehat{EQM}$ $(R^2 + b^2)$	0,012	0,0063
$\widehat{EQM}$ $(R^2 + m^2)$	0,013	0,0063
$\widehat{EQM}$ $(R^2 + c^2)$	0,011	0,0057

2.4.2) $\Delta \cong 2,8$

	$n_1 = n_2 = 20$	$n_1 = n_2 = 40$
$\widehat{EQM}$ (R^2)	0,0078	0,0026
$\widehat{EQM}$ $(R^2 + b^2)$	0,0054	0,0020
$\widehat{EQM}$ $(R^2 + m^2)$	0,0054	0,0019
$\widehat{EQM}$ $(R^2 + c^2)$	0,0066	0,0019

CAPÍTULO 4

CONCLUSÃO DOS RESULTADOS EXPERIMENTAIS

4.1 - INTRODUÇÃO

As conclusões aqui obtidas estão baseadas nos resultados experimentais das tabelas 1 e 2 do capítulo 3. De início compararemos os valores médios dos estimados b_2, c_2 e m_2 com o valor médio de $\hat{B}_2$. Seus resultados não são conclusivos e então, avaliaremos os desvios. Para finalizar nossos estudos, é feita uma comparação dos erros quadráticos médios dos estimadores de P_2; R_2, $R_1 + b_2$, $R_2 + m_2$, $R_2 + c_2$.

As conclusões aqui chegadas, não diferem muito ao que já se conhecem em outros trabalhos. O método bootstrap proposto por Efron [1979], que é não paramétrico, é o mais eficiente já que não faz uso do conhecimento das distribuições populacionais envolvidas no problema.

4.2 - CONCLUSÕES

De início avaliaremos os resultados com base nos valores médios dos estimadores b_2, m_2, c_2 de B_2. Os resultados são variados, dependendo dos valores de Δ , p e n_1, n_2.

Como se percebe na tabela 1, nenhum estimador se demonstra na maioria das vezes, ser o mais próximo do valor médio de $\hat{B}_2$, não obtendo assim diferentes conclusões com respeito à estimação de B_2 somente se baseando nos valores médios de b_2, m_2 e c_2. Para maior clareza, calculamos as diferenças $|\hat{B}_2 - K|$, onde $K = b_2, m_2, c_2$ em todas secções da tabela 1 com respeito aos seus valores médios, como segue abaixo. O lado direito da tabela é a diferença em módulo dos valores médios do estimador bootstrap com o paramétrico.

secção 1.1.1

	$\lvert\hat{B}_2 - b_1\rvert$	$\lvert\hat{B}_2 - m_2\rvert$	$\lvert\hat{B}_2 - c_2\rvert$	$\lvert b_2 - m_2\rvert$
$n_1 = n_2 = 10$	0,01	0,01	0,023	0
$n_1 = n_2 = 20$	0,006	0,005	0,01	0,001

secção 1.1.2

$n_1 = n_2 = 10$	0,009	0,009	0,004	0
$n_1 = n_2 = 20$	0,001	0,000	0,001	0,001

	$\lvert\hat{B}_2 - b_1\rvert$	$\lvert\hat{B}_2 - m_2\rvert$	$\lvert\hat{B}_2 - c_2\rvert$	$\lvert b_2 - m_2\rvert$
secção 1.2.1				
$n_1 = n_2 = 20$	0,001	0,005	0,003	0,004
$n_1 = n_2 = 40$	0,017	0,016	0,017	0,001
secção 1.2.2				
$n_1 = n_2 = 20$	0,002	0,001	0,003	0,001
$n_1 = n_2 = 40$	0,002	0,001	0,001	0,001
secção 1.3.1				
$n_1 = n_2 = 20$	0,001	0,007	0,024	0,006
$n_1 = n_2 = 40$	0,009	0,007	0,001	0,002
secção 1.3.2				
$n_1 = n_2 = 20$	0,006	0,005	0,008	0,001
$n_1 = n_2 = 40$	0,005	0,004	0,002	0,001

	$\lvert\hat{B}_2 - b_1\rvert$	$\lvert\hat{B}_2 - m_2\rvert$	$\lvert\hat{B}_2 - c_2\rvert$	$\lvert b_2 - m_2\rvert$
secção 1.4.1				
$n_1 = n_2 = 20$	0,015	0,007	0,014	0,008
$n_1 = n_2 = 40$	0,010	0,008	0,002	0,002
secção 1.4.2				
$n_1 = n_2 = 20$	0,016	0,019	0,007	0,003
$n_1 = n_2 = 40$	0,005	0,004	0,002	0,001

De acordo com essas diferenças acima, notamos que em algumas situações os valores médios de b_2, m_2 estão mais próximos do valor médio de $\hat{B}_2$ (os resultados de b_2, m_2 são praticamente equivalentes em todas as situações com pequenas diferenças), em relação aos valores médios de c_2. Notamos também que existem alguns casos onde o valor mais próximo de $\hat{B}_2$ é c_2. Os resultados de b_2, c_2, m_2 são mais equivalentes para situaçoes onde o tamanho amostral assume o maior valor (em p = 2; n = 20 e p > 4; n = 40).

O valor médio das diferenças $|b_2 - m_2|$ é 0,002. O estimador bootstrap na maioria das vezes se mostrou bem próximo do estimador paramétrico.

Para p = 2, os valores médios de b_2, m_2 são aproximadamente equivalentes. Os valores de c_2 concorrem com os valores de b_2, m_2, exceto em $\Delta = 1$; $n_1 = n_2 = 10$, onde este atinge a maior diferença. Já em p > 4, a situação não é muito diferente, não temos um estimador que está mais próximo de $\hat{B}_2$ em todas as situações. O c_2 se apresenta em algumas situações mais próximo de $\hat{B}_2$ em relação aos resultados de b_2, m_2, mas geralmente com pequenas diferenças. As maiores diferenças acontecem em p = 6,8.

Portanto, não podemos só nos basear nos resultados médios dos estimadores para chegarmos a uma conclusão com respeito a estimação de B_2.

Faremos agora um estudo com respeito aos desvios, já que a comparação feita entre os valores médios dos estimadores b_2, m_2, c_2 não foi conclusiva.

Verificamos que em todas as situações o desvio de c_2 é bem superior que os desvios de b_2 e m_2. Neste contexto c_2 está atrás de b_2 e m_2. O fato de não avaliarmos agora o desvio de c_2 é que estamos estudando e comparando dois métodos não paramétricos com um paramétrico, e c_2 se mostrou menos eficiente nos desvios. Portanto, as comparações agora são feitas com base na eficiência relativa (ε), já definida no capítulo 3, entre b_2, m_2.

O valores de ε estão compreendidos entre (0,3 - 1,2). Na maioria das situações, os desvios do paramétrico foi inferior ao bootstrap. É claro que nós estamos considerando populações com distribuições conhecidas e, por conseqüência, m_2 leva vantagem. Em cinco casos o valor de ε está próximo de 1, sendo que em 4 deles, este valor é ultrapassado. Isso ocorre nos menores valores de Δ , n_1, n_2 em cada valor de p(2,4,6,8). Portanto, para p fixo, o valor do desvio bootstrap é praticamente equivalente ao desvio do paramétrico onde, Δ e n_1, n_2 assumem menores valores.

Em todas as situações o desvio de c_2 é aproximadamente três a quatro vezes o desvio bootstrap.

A dimensão do espaço p, parece não influir nos resultados experimentais.

Finalmente, verificamos os valores dos erros quadráticos médios (tabela 2). É evidente que em todas as situações o $\widehat{EQM}(R_2)$ foi maior exceto em p = 2, onde se apresenta equivalente aos Erros Quadràticos Médios associados aos estimadores de b_2, m_2,c_2.

Como era de se esperar de acordo com os resultados da tabela 1, os $\widehat{EQM}$ (R_2+ b_2) e $\widehat{EQM}$ (R_2+ m_2) apresentam menores valores, praticamente em todas as situações. Seus resultados são praticamente equivalentes, as diferenças não ultrapassam 0,001.

Portanto, vemos que é de fundamental importância nos problemas que envolvem regras de classificação, estudar o erro de má classificação e o vício, este aparecendo no uso de estimadores da probabilidade de má classificação.

De acordo com que foi visto, nossa conclusão é que o estimador bootstrap parece ser o mais eficiente nos estudos do vício da Razão de Erro Aparente (estimador usado para Razão de Erro Real), visto que na prática é muito difícil encontrar situações como aqui consideradas, isto é, o conhecimento das distribuições populacionais. Caso isso seja possível, não será necessário fazer o uso de estimadores de P_1 e P_2. Eles podem ser calculados diretamente.

É interessante fazer um estudo mais abrangente do que aqui foi desenvolvido, envolvendo outros estimadores de P_1 e P_2 estudados no capítulo 2 e comparando seus vícios com os estimadores no capítulo 3 em um contexto mais geral, considerando várias populações e as matrizes de variâncias e covariâncias diferentes da matriz identidade.

Referências Bibliográficas

1 - ANDERSON, T. W. [1958] - An Introduction to Multivariate Statistical Analysis - *Wiley & Sons*, New York.

2 - ANDERSON, T.W. [1972] - An Asymptotic Expansion of the Distribution of the Studentized Classification Statistic W - *Technical Report 9, Stanford University, to appear Ann. Statist.*

3 - BARTLETT, M. S. [1952] - An Inverse Matriz Adjustment Arising in Discriminant Analysis - *Ann. Math. Stat.*, 22:107.

4 - COCHRAN, W. G. [1968] - Commentary on Estimation of Error Rates in Discriminant Analysis - *Technometrics*, 10(1): 204-205.

5 - EFRON, B. [1979] - Bootstrap Methods, Another Look at the Jackknife - *Annals of Statistics*, 7:1-26.

6 - EFRON, B. [1979] - Computers and the Theory of Statistics - *Siam Review*, 21(4):460-480.

7 - GOELSER, L. [1970] - Técnicas de simulação - *EAESP da FGV*, São Paulo.

8 - JOHNSON, A. RICHARD and WICHERN, D.W. [1982] - Applied Multivariate Statistical Analysis - *Prentice-Hall. Inc., Englewood Cliffs*, New Jersey.

9 - KENNEDY, W. J. and GENTLE, J. E. [1980] - Statistical Computing - *Marcel Dekker*, New York.

10 - KSHIRSAGAR, ANANT M. [1972] - Multivariate Analysis - *Marcel Dekker, Inc.*, New York.

11 - LACHENBRUCH, P. A. and RAY MICKEY [1968] - Estimation of Error Rates in Discriminant Analysis - *Technometrics*, 10:1-11.

12 - MCLACHLAN, G.J. [1976] - The Bias of the Apparent Error Rate in Discriminat Analysis - *Biometrika*, 63:239-244.

13 - MCLACHLAN, G. J. [1980] - The Efficiency of Efron's Bootstrap Approch Applied to Error Rate Estimation in Discriminant Analysis - *J. Statist. Comput. Simul.*, 11: 273-279.

14 - OKAMOTO, MASHASHI [1963] - An Asymptotic Expansion for the Distribution Linear Discriminant Function - *Ann. Math. Statist.*, 34:1286-1301.

15 - SORUM, M. J. [1971] - Estimating the Condicional Probability of Misclassification - *Technometrics*, 13(2):333-343.